THE UNIVERSITY
of EDINBURGH

The Evolution of Restriction-Modification Systems

Edward Bower

Thesis presented for the degree of Doctor of Philosophy

School of Chemistry

The University of Edinburgh

2016

Abstract

Restriction Modification (R-M) systems prevent the invasion of foreign genetic material into bacterial cells and are therefore important in maintaining the integrity of the host genome. The spread of antibiotic resistance, which is proposed to occur via the transfer of foreign genes to the bacterial genome, makes the subject of R-M systems extremely relevant.

R-M systems are currently classified into four types (I to IV) on the basis of differences in composition, target recognition, cofactors and the manner in which they cleave DNA. Kennaway *et al* (2012) proposed that there is an evolutionary link between Types I and II. Comparing the structures of examples from two of the subfamilies of Type II systems (IIB and IIG) to those of Type I structures, similarities can be observed.

Due to the fact that Type II R-M systems cut DNA at fixed positions, they can be used to obtain genetic material selectively. They have therefore proven to be invaluable in molecular biology. One aspect of this project aims to create a novel R-M system, a pseudo-Type II system, by removing the molecular motors from the restriction subunit of a Type I system and fusing the remaining nuclease domain to a known Type I methyltransferase (MTase). This will not only provide evidence to support the theory that evolution has produced a pared down form of the Type I systems in the Type II systems, but it may also become a useful biological tool. This thesis describes the several attempts at doing this and how the subsequent constructs were expressed, purified and assayed to varying degrees of success.

An important characteristic of the Type I systems is their ability to methylate DNA, and it is the mechanism via which host DNA is protected from restriction. This is another subject investigated in this project. As with the nuclease activity of the Type I systems, the site at which DNA is methylated is dictated by the HsdS subunit. It is described here how this subunit can be altered to change the sequence of DNA that is recognised by the system. Again, using Type II system subtypes as a reference, various mutations were made to the HsdS subunit of an MTase from *Staphylococcus aureus*. This is in an effort to bring about a new mode of action, but also to provide further evidence for an evolutionary link between the two system types.

The HsdM and HsdS subunits are expressed from two separate genes at the same locus. There is a frameshift between the genes where the start of the *hsdS* gene occurs a few base pairs upstream from the stop codon of the *hsdM* gene. This work shows that removing this

frameshift creates an MS fusion product, and *in vivo* studies show that this product has methylase activity and can form an active restriction complex when the HsdR subunit is added. The product can also be over-expressed and purified, and shows *in vitro* restriction activity on addition of the HsdR subunit protein.

The HsdS subunit is composed of two target recognition domains (TRDs), each dictating one part of the bipartite recognition sequence. These TRDs can be altered, bringing about a change in the sequence of DNA recognised by the enzyme. In this thesis, it is shown that the C-terminal TRD can be removed and that the subsequent "Half S" enzyme possesses both methylase and restriction activity *in vivo* and that its recognition sequence is different from that of the wild-type enzyme.

After the successful creation of both "MS fusion" and "Half S" recombinant proteins of the Sau1, Type I system from a CC398 strain of *Staphylococcus aureus*, a further construct was produced. This possesses both *in vivo* and *in vitro* activity. The novel "M Half S Fusion" enzyme not only links the two aspects of this project but also creates a structure similar to some seen in the Type II systems. This shows that the Type I systems can be manipulated to change their mode of action but also supports the idea that Types I and II are evolutionarily linked. By making the alterations in a step-wise fashion identifies that these structural changes can create viable enzymes, and that they could have occurred through the process of evolution.

Lay Summary

Restriction enzymes are proteins that cut DNA. They do this by recognising specific sequences of the nucleotides (or bases) that make up DNA, and then catalysing the break of the chemical bonds between them. The opposing action to restriction is modification. Nucleotides can be modified in a number of different ways, but in the case of restriction-modification (R-M) systems, this modification is methylation. A methyl group is a small hydrocarbon group, and this can be added to specific DNA bases. There is a great deal of research into the implications of this relatively small change to DNA, but in this context, the methyl group can prevent restriction at that site on DNA.

R-M systems are a family of enzymes found in bacteria, which carry out both restriction and modification functions. The primary benefit to the bacterial host is that restriction activity serves as a barrier to the invasion of foreign genetic elements, such as viruses (bacteriophage). The recognition sequences of restriction enzymes can occur on viruses, so the enzymes are able to bind to the virus and then degrade it by cutting it into smaller, non-coding fragments. The host's own genetic material is protected from this action by modification. As such, the R-M system is a "cognate" pair of enzymes, which recognise the same DNA sequence and help to preserve the integrity of the genetic material of the host organism. The transfer of genetic material to bacteria is of particular concern, as this is proposed as the reason for the spread of antibiotic resistance. Methicillin-resistant *Staphylococcus aureus* (MRSA) has long been in the public consciousness due to its links to harmful diseases and its resistance to a growing number of antibiotics. The main R-M system in *S. aureus* is the Sau1 family of Type I systems, enzymes from which form the basis of the work shown in this thesis.

In the context of the work presented here, a crucial detail of the R-M system family of enzymes is that they are split into four different types, depending on their structure and mode of action. Perhaps the most important type of these enzymes, in terms of their practical application, is Type II. Restriction enzymes of this type are able to recognise DNA sequences and then cut directly on or around these sites. Given that these sites can be easily identified, these enzymes can be used to manipulate sequences of genetic material selectively. This means that the discovery of these enzymes gave birth to genetic engineering. However, the focus of this project is in fact Type I systems. These are larger, multi-subunit proteins, which carry out both R-M functions in a single enzyme complex. In contrast to Type II systems, Type I enzymes recognise DNA sequences and then translocate (energetically move) the DNA through their

structure, and eventually cut at a random, distant site. The implications of this are that their recognition sequence is not as easily discovered, and they are less useful to biological practice.

What is of key interest, is the relationship between Type I and II R-M systems. A paper published in 2012 by Kennaway *et al*. proposed that structural and mechanical similarities between the two types indicate that they are evolutionarily linked. Sub-types of Type II, IIB and IIG systems, possess both R-M functions in a single enzyme but they lack the characteristic motor function of a Type I. As such, Type IIG and IIB can be thought of as "motor-less" Type I systems. The work presented in this thesis aims to provide evidence for this evolutionary link between the types, by attempting to engineer a pseudo Type II enzyme from the template of a Type I enzyme. By doing this in a step-wise fashion identifies that these structural changes can create viable enzymes, and that they could have occurred through the process of evolution.

Declaration:

I, Edward Bower, hereby certify that this thesis, and the work presented within, is my own. It has not been submitted as partial or complete fulfilment of any other degree or professional qualification.

Signed: Date:

Edward Kenneth Merrick Bower
School of Chemistry
The University of Edinburgh

Acknowledgements:

This work was funded by The School of Chemistry (University of Edinburgh) and the EPSRC.

I have thoroughly enjoyed my time as a Ph.D. student, and have found it to be a challenging but continually rewarding process. For this, I am very thankful to Dr. David Dryden, whose wisdom and kindness has made crucial points of this experience easier. I would also like to thank the members of what was the Dryden group during my studies. I quite literally could not have done my Ph.D. without the expert supervision by Dr. Gareth Roberts and Laurie Cooper. I have greatly appreciated their strong opinions on such things as crisps, beer, and higher education. Their senses of humour have made the dull days pass more swiftly. Dr. John White has also been a great help to me, and has been a fantastic source of knowledge on molecular biology, films and where David is.

I would like to thank everyone at *New England Biolabs*, in particular Dr. Richard Morgan and Yvette Luyten. The resources at NEB are incredible and I feel privileged that I was able to work there.

Another significant influence has been my friend, Dr. Scott Baxter. He has been subjected to hearing the majority of my grievances during my Ph.D. and has always been able to provide both scientific and personal guidance. He is the best scientist I know.

I would also like to thank Dr. Chris Mowat for stepping into the breach and getting things done.

Finally, I would like to thank Professor Dominic Campopiano for allowing me to take this path, and Dr. Jonathan Lowther for making me want to.

List of Abbreviations Used:

3 prime (hydroxyl) end	3’
5 prime (phosphate) end	5’
Absorbance at 600 nm	A_{600}
Acid dissociation constant	pKa
Adenine	A
Adenosine monophosphate	AMP
Adenosine triphosphate	ATP
Alleviation of Restriction of DNA A	ArdA
Amino terminal	N-terminal
Base pairs	bp
Bottom strand	BS
Bovine serum albumin	BSA
*C*5-methylcytosine	m5C
Calf Intestinal Phosphatase	CIP
Carboxy terminal	C-terminal
Clonal cluster	CC
Clustered, regularly interspersed, short palindromic repeat	CRISPR
Community-acquired methicillin-resistant *Staphylococcus aureus*	CA-MRSA
Core variable	CV
CRISPR associated protein	Cas
Cyclic AMP	cAMP
Cytosine	C
Cytosine-guanine base-pair	CpG
Dalton	Da
Deoxynucleoside triphosphate	dNTP
Deoxyribonucleic Acid	DNA
Dideoxynucleoside	ddNTP
Diethylaminoethyl	DEAE
Distilled water	dH_2O
DNA demethyltransferases	dMTases
Double stranded	ds
Double stranded break	DSB
Efficiency of plating	E.O.P.

Endodeoxyribonuclease	ENase
Escherichia coli	*E. coli*
Ethylenediaminetetraacetic acid	EDTA
Genomic DNA	GDNA
Green fluorescent protein	GFP
Guanine	G
Helix-turn-helix	HTH
Hexahistidine tag	HisTag
High-pressure liquid chromatography	HPLC
Homologous recombination	HR
Horizontal Gene Transfer	HGT
Hospital-acquired methicillin-resistant *Staphylococcus aureus*	HA-MRSA
Host specificity for DNA	Hsd
Methylase	M
Specificity	S
Restriction	R
Isoelectric point	pI
Isopropyl-β-D-1-thiogalactopyranoside	IPTG
Kilobase pairs	kb
Kilodalton	kDa
Lambda	λ
Livestock-associated methicillin-resistant *Staphylococcus aureus*	LA-MRSA
Low molecular weight	LMW
Lysogeny broth	LB
Methicillin-resistant *Staphylococcus aureus*	MRSA
Methicillin-sensitive *Staphylococcus aureus*	MSSA
Methyltransferase	MTase
Millivolts	mV
Mobile genetic elements	MGEs
Multilocus sequence typing	MLST
Multiple cloning site	MCS
*N*4-methylcytosine	m4C
*N*6-methyladenine	m6A
New England Biolabs	NEB
Optical density at 600 nm	OD_{600}

Overcome classical restriction	OCR
Percentage weight per volume	(%) W/V
Protein Databank	PDB
Polymerase chain reaction	PCR
Purine	R
Restriction Endonuclease	REase
Restriction-Modification	R-M
Ribonucleic Acid	RNA
Ribosome binding site	RBS
S-adenosylmethionine	SAM
S-adenosylhomocysteine	SAH
Sequence type	ST
Sequencing by synthesis	SBS
Single-molecule real-time	SMRT
Sodium-dodecyl-sulfate polyacrylamide gel electrophoresis	SDS-PAGE
Staphylococcal cassette chromosome	SCC
Staphylococcus aureus bacteremia	SAB
Target recognition domains	TRD
Thymine	T
Top strand	TS
Transcription activator-like effector Nuclease	TALEN
Trichloroacetic acid	TCA
Tris(hydroxymethyl)aminomethane	Tris
Two divalent metal ions	2M
Type I Single polypeptide	Type ISP
Ultraviolet	UV
Ultraviolet/visible	UV/vis
Unmodified	U/M
Unspecified nucleotide	N
Zero-Mode Wavelength	ZMW
Zinc finger nucleases	ZFN

Contents:

Chapter One:

Introduction

1.1. Modification

A genome holds more information than just its sequence of nucleotides. The nucleotide bases can be specifically altered or "modified" in a way that provides further information to cellular machinery. Epigenetics is the study of these other factors, which influence gene expression, regulation and maintenance (Bird 2007). As it is a subject that has been studied quite extensively, a great deal is already known about DNA modifications and their associated enzymes.

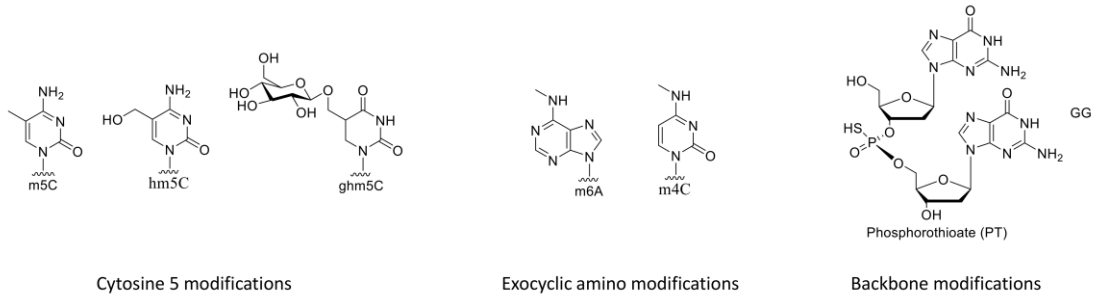


Figure 1: Common DNA modifications (Loenen & Raleigh 2014).

The most common epigenetic modification is methylation, which is facilitated by enzymes known as methyltransferases (MTases) (Furuta et al. 2014). Methylated DNA occurs in both prokaryotes and eukaryotes but seems to involve a more diverse process in prokaryotes (Furuta & Kobayashi 2012; Jeltsch et al. 1999). In Bacteria, there are several different types of DNA MTase that can affect gene expression, and they can do this by methylating the coding or promoter regions of bacterial DNA in a highly specific way (Furuta et al. 2014). Examples of methylated nucleotides that occur frequently in prokaryotes are *C*5-methylcytosine (m5C), *N*4-methylcytosine (m4C) and *N*6-methyladenine (m6A) (Fig. 1). These methyl groups avoid disrupting the base pairs by extending into the major groove of the DNA double helix (Bujnicki 2001). Although they do not affect the secondary structure of DNA, it has been shown that just one of these modifications at a specific target site can regulate the expression of a nearby gene (Furuta & Kobayashi 2012). Studies on bacteriophage have found other methylation modifications. 5-hydroxymethylcytosine (hm5C) and its sugar-attached derivative, 5-glycosylhydroxymethylcytosine (ghm5C), occur as base substitutions in phage DNA. These are therefore not modifications in the traditional sense, as they are not added site-specifically. The phosphate backbone of DNA can also be modified. In prokaryotes, phosphothioester DNA (PT-DNA) can arise, where a sulfur replaces a free oxygen in the phosphate group between the bases (Loenen & Raleigh 2014) This can inhibit both endonuclease and polymerase activities (Xu & Kool 1998).

Of all the DNA modifications, cytosine methylation (mC) is the most well characterised. In fact, m5C is so important to gene regulation, that it is often referred to as the "Fifth nucleotide" (Meng et al. 2015). In mammalian cells, methylation occurs mainly at CpG sites (Furuta & Kobayashi 2012). Therefore, CpG-rich regions of the genome (known as CpG islands) possess a high potential for gene regulation (Kubik & Summerer 2016). In higher eukaryotes, CpG methylation is used to silence genes. It is thought that this occurs when the topology of the DNA changes, due to the increase in hydrophobicity with additional methyl groups (Kaur et al. 2012). Around 60 % of mammalian gene promoter regions consists of CpG islands (Rasool et al. 2015). 70-80 % of mammalian CpG dinucleotides are methylated but this generally does not occur in the promoter regions. The lack of promoter methylation indicates that alterations to the patterns of DNA methylation can have serious consequences on the cell cycle (Meng et al. 2015). In a broader context, DNA methylation has been implicated as the cause of many human diseases, and the number of those that are known is rising (Robertson 2005). The existence and maintenance of DNA methylation is crucial for normal mammalian development and the general fitness of the organism. This has led to an increase in research into controlling epigenetic differences pharmacologically (Robertson 2005).

Most bacterial cells contain restriction endonucleases (REases), which cleave DNA. However, uncontrolled cleavage of DNA would be lethal to the host cell. Therefore, it is often the case that an REase is associated with an MTase, which prevents REase cleavage by methylating the same DNA recognition sequence. An REase paired with its cognate MTase makes up a restriction and modification (R-M) system (Makovets et al. 2004). A popular belief is that methylated nucleotides are too large to fit in the active site of the REase, and as such, the DNA is protected on the basis of steric clashes (Mierzejewska et al. 2016). MTases without a restriction enzyme pair are known as orphan/ solitary MTases (Vasu & Nagaraja 2013). A well characterised example is the DNA-adenine methylation (Dam) methylase, which modifies the adenine in a GATC nucleotide motif (Ratel & Ravanat 2006). Orphan MTases are also sometimes encoded by bacteriophage as an anti-restriction measure. The MTase coded by *Bacillus* phage H2 shares a recognition sequence with the bacterial hosts REase, BamHI. The phage is therefore able to evade restriction by this enzyme (Wilson & Murray 1991). All of the methylation modifications protect the DNA from being cleaved by the conventional REases. However, for each modification, there is also at least one enzyme that attacks DNA only when the modification is present (Loenen & Raleigh 2014). R-M enzymes that possess this characteristic fall into the IIM or Type IV categories. Type IV enzymes differ from Type IIM, due to their lack of a defined recognition sequence (Roberts et al. 2003).

There is a proposed mechanism for the methylation of the *C*5 of cytosine, and another distinct mechanism for the amino MTases, which methylate the *N*4 of cytosine or the *N*6 of adenine. Amino methylation probably occurs via the deprotonation of the exocyclic nitrogen target, and to it the direct transfer of the methyl group. The *N*6 is activated when it is polarised by the formation of hydrogen bonds between the hydrogens in the amino group and the surrounding active site residues (Bheemanaik et al. 2006) (Fig. 2A and B). The active site of m6A MTases consists of a well conserved (D/N)PP(Y/F) motif, whilst the majority of the m4C MTases possess an SPP(Y/F) sequence (Jeltsch et al. 1999). It has been found that some enzymes (such as M.EcoRI) can methylate *N*4 of cytosine, despite their target being *N*6 adenine. This low specificity in the amino MTases and their shared exocyclic NH_2 target, provides evidence that the methylation of *N*4 cytosine and *N*6 adenine proceed via the same mechanism (Jeltsch et al. 1999).

The proposed mechanism for the methylation of the cytosine pyrimidine ring requires an enzyme-bound intermediate (Fig. 2C). The cysteine thiol acts as a nucleophile and forms a covalent link to the *C*6 of the cytosine. This activates the *C*5 for the addition of the methyl group, donated by SAM (Scavetta et al. 2000). The activation of the *C*5 by the enzyme is necessary as it is not otherwise a strong nucleophile. The enzyme nucleophile is eliminated when the *C*5 deprotonates (Zangi et al. 2010). Several other amino acids play key roles in this mechanism. Alignments of MTases from different sources highlighted a number of conserved motifs. A glutamic acid, valine and two arginines (Glu, Val, Arg, Arg), along with the cysteine are the five amino acids that make up the enzyme active site. It is thought that the Glu and two Arg residues allow the nucleophilic attack by the cysteine, whilst the Val stabilises the cytosine (Zangi et al. 2010). Much about this mechanism has been elucidated by examining the structure of M.HhaI, an MTase from the bacterium *Haemophilus haemolyticus.* The crystal structure of this protein showed it bound to its DNA substrate and the SAM cofactor (Shieh & Reich 2007). The structure of the SAM-enzyme complex appears to have two conformations, dictated by a large loop formed by nineteen residues. In the presence of a non-specific DNA sequence, the loop is open but closes when the enzyme is bound to its target sequence (Estabrook et al. 2004).

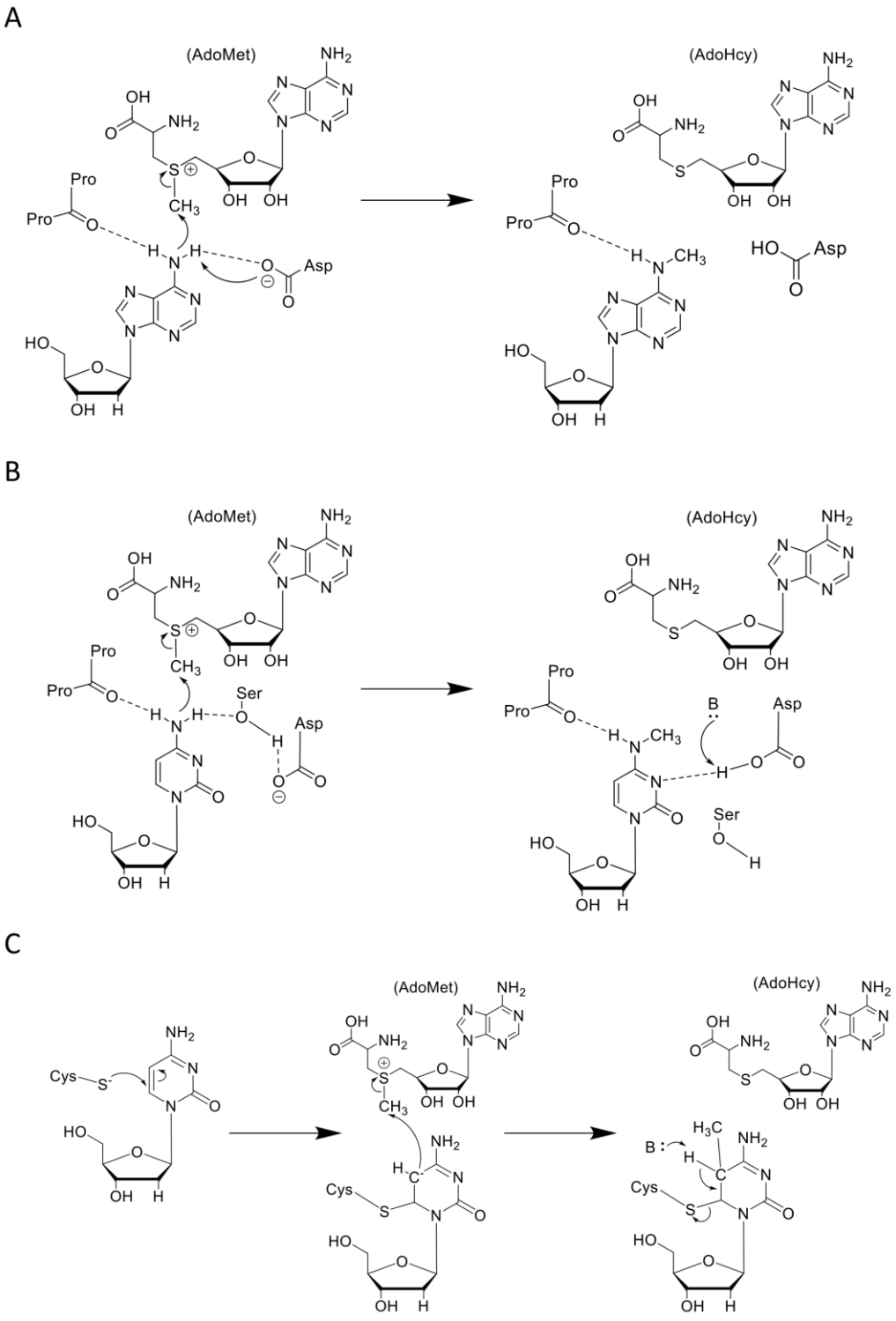


Figure 2: The proposed mechanisms for adenine methylation (A) (Scavetta et al. 2000), *N*4 cytosine methylation (B) (Gong et al. 1997), and *C*5 cytosine methylation (C) (Wu & Santi 1987).

There is a conserved structural motif amongst the MTases, which consists of a seven stranded β-sheet core known as the "SAM-dependent MTase fold" (Cheng & Roberts 2001). Questions are raised as to how this structure can facilitate the methylation of such a diverse range of substrates, like RNA, protein and small molecules. In particular, it was hard to imagine how it could act on large proteins as well as specific DNA bases, which would otherwise appear to be inaccessible. The answer was given by the M.HhaI structure, which had turned its cytosine substrate 180^{o} out of the DNA helix (Shieh & Reich 2007). This process by which the nucleotide is extra-helically bound is known as base flipping. It allows access to target bases and enables the enzyme to bind the DNA tightly (Estabrook et al. 2004). Substituting valine with alanine inhibits base flipping and enzyme activity. The DNA binding affinity of the enzyme also decreases dramatically, but can be recovered using a DNA substrate that does not contain the target cytosine. This confirms that base flipping is necessary for catalysis and that it is stabilised by the valine residue (Estabrook et al. 2004). The high structural conservation between the *C*5 MTases suggests that base flipping is a common mechanism, and has been shown to occur in several other systems (Roberts & Cheng 1998). Human uracil DNA glycosylase, which removes uracil from double or single stranded DNA has also been found to use base flipping to facilitate its reaction (Poole et al. 2001). Despite this phenomenon being observed in a m5C MTase, it is also the most likely mechanism in the amino MTases. All of the known MTases show a higher binding affinity for their recognition sequence if their target base is in a mismatch pair. It is assumed that this eases base flipping due to the weakened Watson-Crick bonds between the bases (Jeltsch et al. 1999). It is not yet known precisely how base flipping occurs, but there are two different suggestions. One proposes that flipping occurs via three steps, where specific residues in the protein are inserted into the DNA helix, these then lengthen the distance between the phosphates surrounding the target base. The enzyme then pulls the base out and into its active site, where it remains for the duration of the reaction. The opposing view is that the DNA helix is a dynamic structure, and the bases flip out as part of its normal movement. It is thought that the enzymes recognise and take advantage of this transient state. Structural and biochemical evidence from M.HhaI and T4 endonuclease V suggests that the three-step mechanism is most likely (Roberts & Cheng 1998).

In bacteria, m5C and m4C are considered responsible mainly for protection against restriction, whilst m6A plays a part in several different processes. These include DNA replication and repair, and control of virulence (Ratel & Ravanat 2006). It was long thought that m6A did not occur at all in eukaryotes, but it has been discovered that not only does it exist in higher eukaryotes, such as insects and algae, but that it has a key role in gene regulation (Sun et al.

2015). That this modification has been undiscovered until recently is most likely due to the lack of sensitivity in any of the techniques used to analyse the DNA (Ratel & Ravanat 2006). The poor sensitivity is not the only reason quantification of methylated bases is difficult. It is also an inherent structural issue of m5C, which can lead to mutations in the genome. m5C can deaminate to become thymine, and cytosine to uracil, and this can happen spontaneously (Fig. 3). It is the responsibility of uracil DNA glycosylase to repair this mutation (Poole et al. 2001). m6A MTases can also be encoded by viruses and have been found in bacteriophages T4 and Mx8, and archeal viruses ϕCH1 and SNDV. This implies that m6A is involved in viral infection, and highlights the ubiquity, and by extension the importance of m6A, which is becoming known as the 6th base (Ratel & Ravanat 2006).

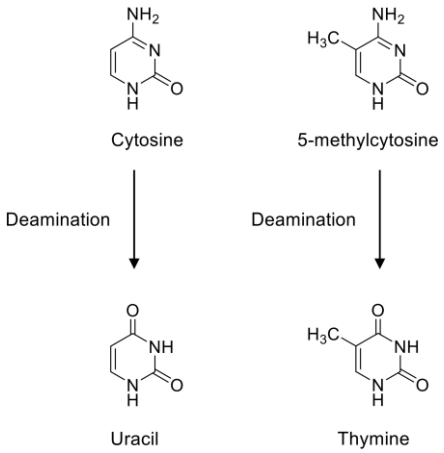


Figure 3: The spontaneous change from cytosine to uracil and m5C to thymine (Poole et al. 2001).

The most frequently used biological methyl donor is a sulfonium compound called S-adenosylmethionine (SAM or AdoMet), which is also the second most common enzyme substrate after ATP (Salyan et al. 2006; Fontecave et al. 2004). SAM is made when methionine and ATP are joined in a stereospecific reaction, catalysed by SAM synthetase (or methionine adenosyltransferase). This biosynthesis only produces an *S*-configuration at the positively charged sulfur atom (Fontecave et al. 2004). The methylthiol of the methionine would otherwise be relatively inert but the charged sulfur makes it thermodynamically unstable, making it far more reactive. There are a number of different atoms to which SAM can transfer its methyl group. These include sulfur, nitrogen and oxygen, which are polarisable nucleophiles, and carbon, in the form of a carbocation (Cheng & Roberts 2001). MTases transfer a methyl group from SAM to a number of different substrates, including proteins, RNA and DNA (Salyan et al. 2006). Interestingly, there is very little sequence homology

between the SAM-dependent MTases. They do however share some structural homology. The SAM cofactor is often bound by an α/β/α sandwich structure, which provides specific contacts to the SAM and the enzyme substrate (Salyan et al. 2006). This substrate accepts the methyl group from SAM, via an S_N2 mechanism (Figure 2). The reaction proceeds by the methyl acceptor acting as a nucleophile, whilst the SAM is very electrophilic (Fontecave et al. 2004). The methyl group is donated, and the SAM is converted to S-adenosylhomocysteine (SAH or AdoHcy). This product can be hydrolysed to adenosine and homocysteine by SAH hydrolase (Fontecave et al. 2004). All MTases possess an FXGXG amino acid motif, or variant thereof, which coordinates with the SAM cofactor (Jeltsch et al. 1999).

DNA in the cell consists of two strands of nucleotides wound together to form a *B*-form double helix. In this state, a recognition sequence of an MTase can occur on either side of the helix. This leads to a distinction between levels of methylation, where both sides can be methylated in fully methylated DNA or only one strand in hemimethylated. The semi-conservative replication of fully methylated DNA results in hemimethylated DNA. Given the necessity to protect host DNA, hemimethylation is enough to prevent restriction (Wilson & Murray 1991). Additionally, R-M system MTases strongly prefer hemimethylated DNA as a substrate (Mierzejewska et al. 2016). These MTases can modify non-methylated DNA, but at a slower rate. This leads to the distinction between host and foreign DNA (Powell & Murray 1995). Adenine methylation is also crucial to the post-replicative mismatch repair system. The difference in strand methylation produced after replication is used to identify the daughter strand, and allows mismatched bases to be changed (Horton et al. 2006).

The other aspect to the pattern of methylation is its effect on cellular processes. As such, MTases are responsible for establishing and continuing the correct distribution of methyl groups on DNA. In eukaryotes, MTases are divided into different groups, depending on the pre-existing state of their DNA substrate. Maintenance MTases act on hemimethylated DNA, and therefore regulate methylation after DNA replication by adding a methyl group to the newly synthesised strand. *De novo* MTases on the other hand, act on either hemi- or non-methylated DNA, and can establish a new pattern of methylation (Piccolo & Fisher 2014). The control of these gene signals plays a key role in the response of higher organisms to their environment and stage of development. Given that it is critical, DNA methylation, even *de novo* methylation, was thought to be inherited and stable. Breaking the C-C bond between *C*5 and the methyl group would take a high reaction energy, and so was considered not thermodynamically possible (Ramchandani et al. 1999). Nevertheless, it was acknowledged

that changes had to occur throughout the lifetime of an organism, and so it was proposed to occur via base excision or other indirect pathways. Surprisingly, it was discovered that DNA methylation can actually be reversed in higher organisms. DNA demethyltransferases (dMTases) remove the methyl group from m5C, in a reaction that produces methanol (Ramchandani et al. 1999). dMTases act site-specifically on CpG dinucleotides in fully or hemimethylated DNA, in a process known as active demethylation. Demethylation can also occur passively when the expression of MTases is halted or the enzymes themselves are inhibited (Piccolo & Fisher 2014).

1.2. SMRT Sequencing (*Pacific Biosciences*)

DNA sequencing provides a surfeit of information and has been vital to the better understanding of biological systems (Korlach & Turner 2012). Since its introduction in the 1970s, Sanger sequencing has been the dominant technique for obtaining the sequence of the four canonical bases in genes. It provided an automated, high-throughput method, which used analogues of the deoxynucleoside triphosphates (dNTPs) to terminate the DNA chain. The analogues 2', 3'-dideoxynucleoside (ddNTP) and arabinonucleoside inhibit DNA polymerase, and therefore stop the further addition of bases to the DNA molecule (Sanger et al. 1977). The low error rate of DNA polymerases has meant that this method has been a mainstay for decades. However, Sanger sequencing does not utilise the high catalytic rate or turnover of the polymerase (Eid et al. 2009). Furthermore, the study of epigenetics makes it necessary to go beyond the four bases and identify their modifications.

Recently, a way of determining large sequences of DNA and detecting base methylation has come into practice. This can be done relatively quickly, with very little labour (Korlach & Turner 2012). Single-molecule real-time (SMRT) sequencing from *Pacific Biosciences* is an significant advance in the new generation of sequencing techniques (R. J. Roberts et al. 2013). The different technologies from *Illumina, 454* and *Ion Torrent* seem to be limited by short read lengths and amplification bias. There are several benefits to increasing read length. These include a decrease to the duration of the procedure and the subsequent sequence assembly, reducing cost and increasing accuracy. Other next-generation technologies have achieved this but as their methods involve regulating the activity of the enzyme, they have not been able to produce sequences above 400 bp (Eid et al. 2009). SMRT sequencing on the other hand, can read much longer sequences with incredible accuracy, and identify DNA modifications. This is an extremely powerful tool in the study of DNA MTases, as it can identify DNA recognition sequences (Murray et al. 2012; R. J. Roberts et al. 2013). This is particularly useful when it comes to studying R-M systems. which possess a nuclease that does not cleave DNA at a fixed position, and do not therefore have an easily identified recognition sequence (Clark et al. 2012).

The Sanger method uses ddNTP analogues of each of the bases, in a mixture with the canonical base. The analogues terminate the chain of transcription wherever the normal base would be incorporated, in a process known as sequencing by synthesis (SBS) (Sanger et al. 1977; Shendure et al. 2011). Originally, the ddNTPs were labelled with P^{32} but this has now been replaced with fluorescent-tagged dideoxynucleotides, resolved temporally (Prober et al. 1987).

Despite this quite significant improvement to the technique, a compromise still has to be made between turnover and read length (Shendure et al. 2011). SMRT sequencing is similar in many ways. It too uses fluorescence-tagged nucleotides in an SBS reaction using DNA polymerase. In contrast though, SMRT is conducted in "real-time", which means that DNA polymerisation can actually be observed directly, with base pair resolution. The benefit of using "single molecules" in SMRT is that the signals only end when the polymerase dissociates from the template (Eid et al. 2009). The result is an average read length of around 3000 bp, with a maximum of over 20,000 bp (R. J. Roberts et al. 2013).

The novel chemistry used by SMRT is a key part of the real-time process. The normal nucleotides are replaced entirely with labelled dNTPs. The fluorophore label is phospholinked to the end of the nucleotide, and leaves when the dNTP is incorporated by DNA polymerase into the growing DNA chain. Each dNTP possesses a fluorophore with a distinct emission wavelength, which is detected in what is known as the Zero-Mode Wavelength (ZMW) region when the nucleotide is being added. The phospholinked dNTPs are A555-dATP, A568-dTTP, A647-dGTP and A660-dCTP. None of these inhibits the polymerase, allowing close to normal reaction kinetics. The period of fluorescence ends when the tag leaves the ZMW, after the formation of the phosphodiester bonds cleaves the tag from the nucleotide. A new period begins with the next tagged nucleotide (Eid et al. 2009). This means that the period of fluorescence is a measure of polymerase turnover, whilst the colour emitted is specific to each tag (Flusberg et al. 2010). It has been shown that the secondary structure of the template DNA has an effect on enzyme kinetics, and so too does methylation. Put simply, methylated bases slow the action of the polymerase and result in a longer duration of fluorescence relative to a non-methylated control (Fig. 4).

There are two potential problems with the SMRT technique. The first is that perhaps not all bases will be methylated at each specific site in a genomic sample, whilst the other is an 11 to 14 % error in individual reads. However, both of these issues are resolved by circular consensus sequencing (R. J. Roberts et al. 2013). This involves the reading of individual molecules multiple times to provide an average (Flusberg et al. 2010). This also significantly improves the accuracy of the nucleotide sequence read (R. J. Roberts et al. 2013).

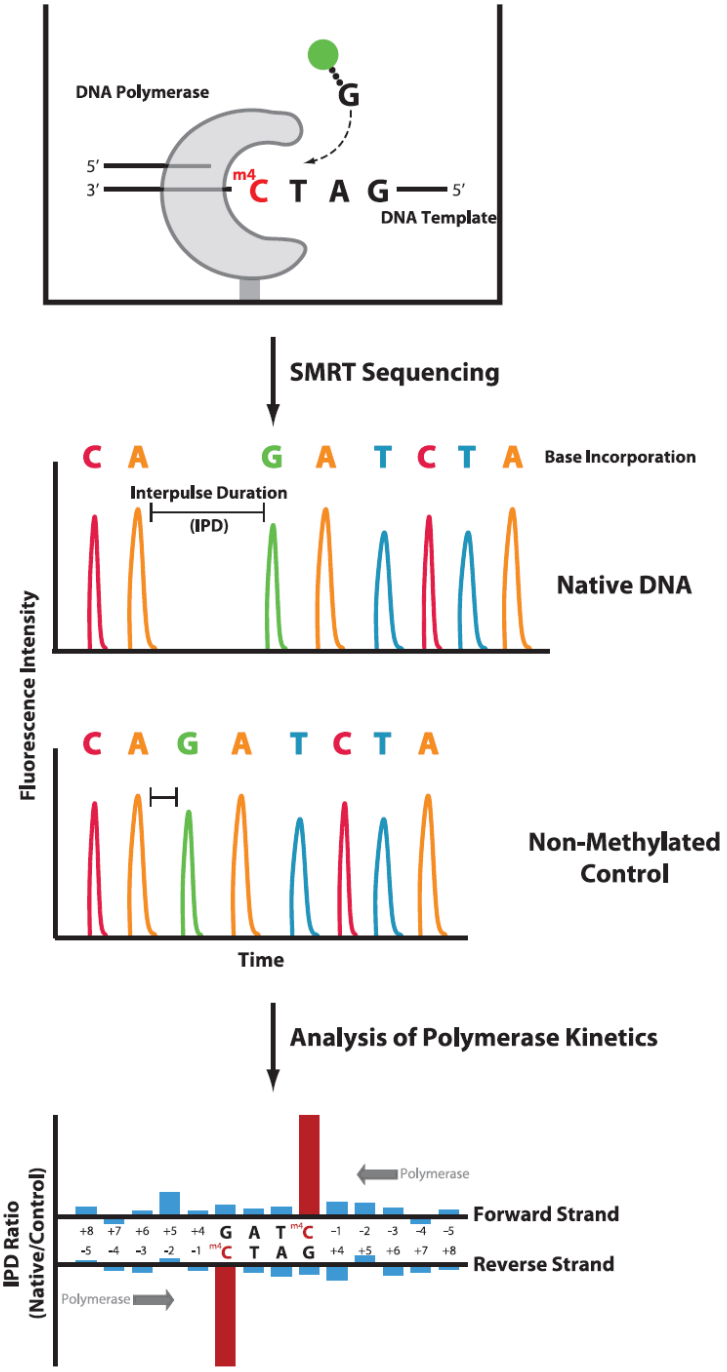


Figure 4: Method of sequencing DNA and identifying modification using SMRT technology (Clark et al. 2012).

1.3. DNA Cleavage

The cleavage of DNA is integral to the regulation of a multitude of cellular processes (Gowda et al. 2014). These include unwinding DNA to facilitate replication, to initiate recombination, and as a barrier to mobile genetic elements (MGEs) in prokaryotes (Yang 2011; Jurėnaitė-urbanavičienė et al. 2016). Breaks in the DNA chain can occur in many different ways. These range from methods controlled by the cells themselves, to an effect of external forces such as radiation (Schulte-Frohlinde 1987). Different techniques to cleave DNA experimentally are also wide-spread, and can provide a wealth of information on DNA structure and binding (Tsen & Levene 2004). Breaks to the nucleotide chain can be made intentionally, using synthetic chemical agents, such as ferrous EDTA or uranyl acetate, or by biological agents like bleomycin (Sigman & Chen 1990). Perhaps most important though, was the discovery of enzymes that cut DNA at specific, predetermined sites, allowing the deliberate selection and excision of genes from a genome (Saravanan et al. 2016). This makes them invaluable biological tools (Bickle & Kruger 1993).

Enzymes that cut DNA are known as nucleases, and these are separated into groups depending on their substrate and mode of action (Yang 2011). Single or double stranded (ds) DNA can be cut at any site within it, by endonucleases, or one nucleotide at a time from its ends, by exonucleases (Kushner 1974). Some nucleases bind and cleave RNA and are vitally important to gene expression (Abelson et al. 1998). Some provide other enzymes access to the DNA, for example the 3' to 5' exonuclease that facilitates proof-reading. Other nucleases cut DNA in order to change its tertiary structure. Topoisomerases are a family of enzymes that create a temporary break in one or both strands of DNA (Liu & Wang 2016). Enzymes in the Topoisomerase I subfamily break one stand, causing it to unwind around the other, whilst Topoisomerase II enzymes create a double stranded break (DSB), allowing another double helix to pass through (Champoux 2001).

Nucleases are also involved in apoptosis (Yang 2011). Apoptosis is the name given to programmed cell death, or the regulated killing of cells by the host organism. It is important in the formation of correct structures during development, the continual renewal of blood cells and the disposal of damaged cells (Parrish & Xue 2006; Thompson 1995). The disruption of such a significant process also has serious implications, and is linked to several diseases, like cancer and neurodegenerative disorders (Thompson 1995). One of the steps in apoptosis is the cleavage of chromosomal DNA into fragments of around 180 bp. This prevents the cell from further replication, and passing on its genetic material. In mammals, one of the nucleases

responsible for this degradation of DNA is the 40-kd DNA fragmentation factor (DFF40) or Caspase-activated deoxyribonuclease (CAD) (Parrish & Xue 2006).

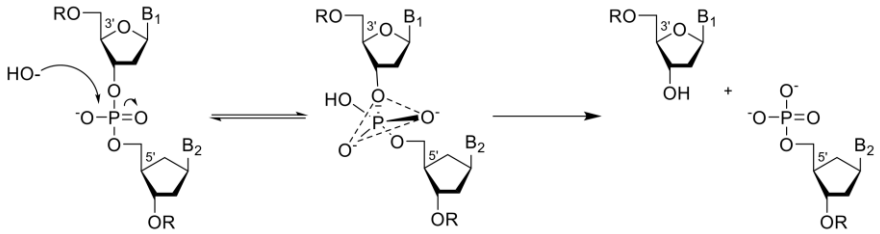


Figure 5: Proposed mechanism for the cleavage of DNA
by the hydrolysis of its phosphodiester bonds (Gowda et al. 2014).

DNA hydrolysis is the path by which DNA is cleaved enzymatically. This is where the phosphodiester bonds are broken by the enzyme in the presence of water (Gowda et al. 2014). Due to the extremely high stability of DNA in water, an enzyme is required to speed up the hydrolysis reaction by over 10^3-fold, to bring the reaction time to within a couple of minutes (Wolfenden & Snider 2001). The metal ion cofactors of the enzyme activate water by acting as Lewis acids, which initiates the first step of the proposed S_N2 mechanism. It proceeds with the nucleophilic attack of the phosphate group between the bases by the subsequent hydroxide ion. This leads to the formation of a five-coordinate intermediate, which is stabilised by the enzyme. The collapse of the intermediate results in the separation of the bases, one of which leaves as an alcohol (Fig. 5) (Gowda et al. 2014).

The two divalent metal ions (2M) bound by the enzyme, are crucially important nuclease cofactors (Yang et al. 2006). In most metallonucleases these are Mg^{2+} ions, are on the phosphate backbone side of the chain, and are 3.8 to 4 Å apart, either side of the scissile phosphate (Steitz & Steitz 1993; Palermo et al. 2015). A theory put forward in 1993 suggested that the two ions have separate mechanistic responsibilities (Yang et al. 2006). Both ions activate water for nucleophilic attack but they do so with complimentary and distinct actions. One ion lowers the pKa of the water molecule and binds the hydroxide ion to allow the attack of the target phosphate (Yang et al. 2006). The second ion enables the exit of the oxygen, whilst both ions then play a role in stabilising the subsequent five-coordinated transition state (Steitz & Steitz 1993). After this, it is the second ion that destabilises the intermediate and facilitates the formation of the products (Yang et al. 2006). Studies of the metal binding sites of Klenow fragment have shown that this second ion is not present when no DNA is bound. This suggests that the second ion binds the enzyme after the substrate (Sträter et al. 1996).

Although since validated, the 2M theory was under some debate. It was shown that some mutated enzymes remain active, despite lacking the ability to bind the first metal ion (Palermo et al. 2015; Yang et al. 2006). On the other hand, mutants lacking the ability to bind the second ion lose activity almost completely. This highlights the importance of stabilising the reaction intermediate (Sträter et al. 1996). There was also a dispute as to how many metal ions were required for catalysis (Yang et al. 2006). Several more structures have been found that corroborate the 2M theory however, and have also provided further information. The cations are bound by a motif of acidic residues, conserved in the metallonucleases. The "DEDD" motif chelates the metal ions into the enzyme active site and directs them towards the DNA substrate. The carboxylate side chains of each residue can bind one or both ions, and are aided by the slight negative charge of nearby oxygen atoms from water molecules (Palermo et al. 2015). Surprisingly, despite the charged nature of their acidic side chains, the metal-binding residues appear to have no role in stabilising the reaction directly (Sträter et al. 1996). Computational studies by Palermo *et al* have investigated the effect of a third ion in the active site of metallonucleases. They found that in RNase H, the presence of a third ion does not affect activity. However, at concentrations of 50 mM and above, the third Mg^{2+} is closer to the site of catalysis, moving the active water molecule away from its optimal location (Palermo et al. 2015). This mostly results in enzyme inhibition but does not appear to be the case for all of the metallonucleases. Several of the enzymes possess three metal ion binding sites, which have been shown to be required for catalysis. They also still adhere to the general 3- 4 Å spacing of their closest two cations (Sträter et al. 1996). Interestingly, EcoRV also has three ion-binding pockets but its reaction appears to be coordinated by only two ions. It has therefore been proposed that the cations are mobile and move positions during the reaction (Dupureur 2008). Whether there is a role for a third ion in the metallonucleases is uncertain, but it has been suggested that it could be to allow the reaction products to leave the active site (Palermo et al. 2015).

Mg^{2+} is the most commonly used metal ion cofactor in metalloenzymes, due to its ability to bind six ligands with octahedral geometry. It also maintains a uniform coordination distance of 2.1 Å. (Yang et al. 2006). The Ca^{2+} divalent cation has a relatively large atomic radius and can coordinate up to nine ligands, with various geometries. The accommodating nature of Ca^{2+} makes it less appropriate for hydrolysis of phosphodiester bonds, as it is less able to destabilise the enzyme-substrate complex (Yang et al. 2006). The metal cation Mn^{2+} has similar properties to Mg^{2+}, and is therefore an adequate alternative. However, the cellular concentration of Mg^{2+} is much higher than Mn^{2+}, and it is therefore a more likely enzyme cofactor. Interestingly,

Mn^{2+} can exist in several different coordination states, of various different lengths. As such, it continues to facilitate catalysis, despite mutations of the catalytic residues and changes to substrate sequence. (Yang et al. 2006). Zn^{2+} ions are also common nuclease cofactors (Yang et al. 2006). Zn^{2+} can accommodate octahedral or tetrahedral geometries and binds ligands with higher affinity than Mg^{2+}, creating complexes with higher stability (Dudev & Lim 2014). This means that Zn^{2+} can replace the Mg^{2+} cofactor in some proteins and inhibit others. The relative concentrations of these metals in the cell is again what selects Mg^{2+} for most metallonucleases (Dudev & Lim 2014).

Despite lower cellular concentrations, Zn^{2+} is still found in many metalloenzymes. A pertinent example of this is zinc fingers (Dudev & Lim 2014). Zinc fingers were first described in 1985, as a repeating motif of the 40 kDa Transcription factor IIA protein (Klug 2010). Biochemical characterisation of this protein showed that the motif consisted of nine independent 3 kDa sections, each with a strong affinity for DNA. Each 25-amino-acid unit was separated by a linker of 5 amino acids and stabilised by a Zn^{2+} ion, bound in a tetrahedral conformation by a Cys-Cys-His-His consensus sequence (Miller et al. 1985). These sections were later referred to as zinc fingers, due to the manner in which they protrude from the main protein structure, and the way they "grip" DNA (Klug 2010). Also conserved were three large hydrophobic amino acids. A repeating pattern of leucine, phenylalanine and tyrosine residues are proposed to stabilise the structure by creating a large hydrophobic region (Klug 2010). Each consecutive "finger" binds a sequential DNA triplet, and does so via the side chains of only three specific residues (Klug 2010; Pavletich & Pabo 1991). This discovery led to the suggestion that the DNA recognition sequences could be easily altered with the mutation of these key residues (Carroll 2011).

FokI is a Type IIS restriction enzyme from the bacterium *Flavobacterium okeanokoites* (Wah et al. 1998). Type IIS systems cleave DNA at a non-uniform position to one side of their recognition site (Halford et al. 2011). FokI cleaves the 5' to 3' DNA strand 9 bp downstream from its 5'-GGATG-3' recognition sequence, whilst the 3' to 5' strand is cleaved 13 bp away (Carroll 2011). Interestingly, although FokI is a monomer in solution, it is not active in this state. A single unit of this protein will bind to the 3' to 5' strand but shows no nuclease activity. FokI dimerises when a second monomer binds the 5' to 3' strand, and the enzyme becomes an active nuclease. The FokI monomer consists of two separate domains. Its C-terminal carries the nuclease, whilst its N-terminus is responsible for DNA recognition (Halford et al. 2011). This domain is made up of three separate subdomains (D1, 2 and 3), all of which possess a

helix-turn-helix (HTH) motif. The symmetry of the HTH often corresponds to a palindromic DNA substrate, although the HTH is found in most sequence-specific DNA binding proteins (Klug 2010; Pabo & Sauer 1984). Remarkably, the recognition of specific bases can be linked directly to these subdomains, where D1 binds GGATG and D2 binds GGATG. D3 appears not to make contact with the nucleotides, but with the second FokI monomer (Pingoud & Jeltsch 2001). Li *et al* found that in the presence of its substrate, trypsin digestion separates FokI into 41 kDa and 25 kDa peptides. The 25 kDa fragment was found to cleave both methylated and non-methylated DNA non-specifically (Li et al. 1992). This led to the realisation that the nuclease domain could be separated from the rest of the protein and attached to other DNA-binding domains, imparting new specificity (Carroll 2011). By joining zinc fingers to the nuclease domain of FokI, a new method for targeting and editing genomic DNA was born (Carroll 2011; Swarthout et al. 2011). Zinc finger nucleases (ZFNs) provide an adaptable way of creating a site-specific DSB in DNA, which can be used to target and replace genes in a genome (Fig. 6A). Homologous recombination (HR) is the way in which a cell maintains genome integrity after a DSB, but it can be used for the modification of a desired gene (Durai et al. 2005). Therefore, creating a specific DSB with a ZFN can make changes to the genome by enlisting the cells own machinery for DNA repair (Swarthout et al. 2011). However, Halford *et al* argues that as a tool for genome therapy, ZFNs might not be as powerful as hoped. Despite the necessity for dimerisation, the two parts of the recognition site may only be close to each other in space but not on the DNA chain. Therefore, ZFNs often produce off-site cleavage (Halford et al. 2011). Whether or not ZFNs are a viable tool for gene therapy is still up for debate. In principle however, they prove that nucleases can be manufactured using the DNA cleavage domain of FokI.

Transcription activator-like effectors (TALEs) from the bacterial plant pathogen, *Xanthomonas campestris*, were discovered in 1989 (Boch et al. 2009). TALE proteins are important virulence factors, which are secreted into the plant cell cytoplasm and change transcription of the host genome by behaving like eukaryotic transcription factors. Most often, TALEs promote transcription of genes, whose products make the plant more vulnerable to the bacterium (Christian et al. 2010). The protein possesses three separate domains, for transcription activation, nuclear localisation, and a region of tandem repeats that binds DNA (Boch et al. 2009). Each repeat consists of 30 to 35 residues but only corresponds to a single nucleotide (Chandrasegaran & Carroll 2016). Importantly, only two amino acids (at positions 12 and 13 of the repeat unit) correspond to the identity of the bound nucleotide (Ain et al. 2015). Therefore, these can be altered in order to engineer a new recognition sequence (Boch

et al. 2009). Continuing from the work on ZFNs, TALE proteins can also be used to create another way to cut DNA *in vivo*. Transcription activator-like effector nucleases (TALENs) are the combination of TALE repeats and the cleavage domain of FokI (Ain et al. 2015). TALENS can produce specific DSBs in genomic DNA, but given the off-site cleavages seen using ZFNs, do so with a lower level of cytotoxicity (Fig. 6B) (Chandrasegaran & Carroll 2016). A significant drawback to this technique is in the design and production of new specificities. The several amino acid repeats in the structure make it difficult to avoid mis-priming during PCR and therefore, special cloning techniques have to be employed (Ain et al. 2015).

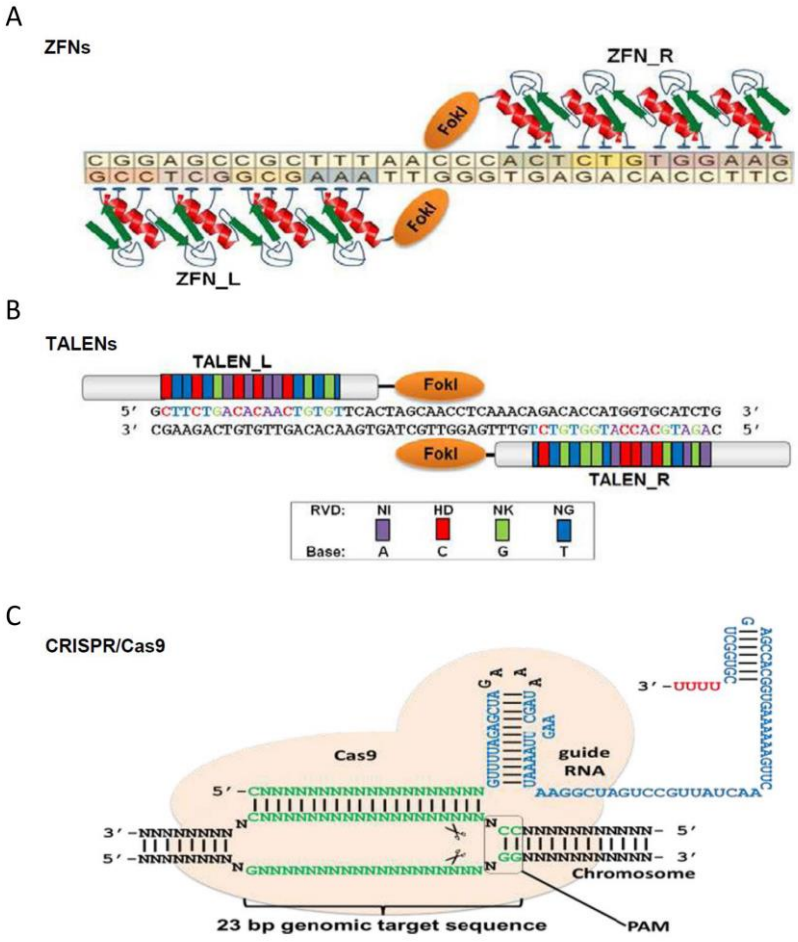


Figure 6: Diagram to show the mode of action of the three nucleases used for genome engineering. Zinc finger nucleases (A) Transcription activator-like effector nucleases (B) and CRIPR-Cas9 (C). Adapted from Chandrasegaran & Carroll 2016.

ZFNs and TALENs were significant advances in the effort to create a tool for genome modification, but are not without their problems. Utilising the clustered, regularly interspersed, short palindromic repeat (CRISPR) and the CRISPR associated protein 9 (Cas9) system has revolutionised the field, by providing an easy way to edit genomes (Doudna & Charpentier

2014). Prokaryotes use CRISPR-Cas9 as an acquired and adaptive immune response against foreign genetic elements (Ceasar et al. 2016). The CRISPR loci contain short nucleotide repeats, with intercalated sequences derived from plasmids and phage (Doudna & Charpentier 2014). From the CRISPR sequences are transcribed CRISPR RNA (crRNA), which anneal to trans-activating RNAs (tracrRNAs). This RNA duplex is bound to the Cas nuclease, and used to guide it to a target sequence (protospacer) in invading viruses, via complementary base-pairing (Doudna & Charpentier 2014). The Cas protein creates a DSB in the foreign genetic material at a site 3 bp away from the Protospacer Adjacent Motif (PAM), which results in the degradation of the foreign element. In the case of the most widely used Cas protein (from *Streptococcus pyogenes*), the PAM is an NGG sequence (Ceasar et al. 2016) (Fig. 6C). A significant breakthrough came when it was discovered that the crRNA/tracrRNA duplex could be manufactured as a single guide RNA (sgRNA), whilst maintaining its ability to bind both the target DNA and the Cas protein (Doudna & Charpentier 2014). Therefore, engineered sgRNAs can be used to direct Cas9 to create DSBs at desired sites. Designing custom sgRNAs is easy compared to the protein engineering involved in the ZFN and TALEN techniques, and results in the same targeting efficiency (Chandrasegaran & Carroll 2016).

1.4. Restriction-Modification Systems and Bacteriophage λ

Found exclusively in single-celled organisms and their viruses (phage), restriction-modification systems are a prokaryotic organism's foremost defence against invading foreign DNA (Bujnicki 2001). Although thought to be a "primitive" prokaryotic immune system, the ubiquity of R-M systems indicates their importance (Vasu & Nagaraja 2013). In the early 1950s, it was observed that the survival of phage was affected by the organism that it had last infected. Phage that had successfully infected a specific bacterial strain could then spread to other cultures of the same strain. Conversely, it was not very common for phage to proliferate in different strains. This was the point at which the notion of modification was first conceived; the bacterial strain had left an imprint on the phage such that it could propagate in the same strain. It was later discovered that this "imprint" was a DNA modification by methylation (Murray 2000). It was also shown that without this modification, R-M systems would recognise the phage as foreign and it would therefore be subject to degradation, which in most cases produced specific fragments of DNA. It is due to this characteristic that the term "restriction" was used for these enzymes, as they restrict the growth of the phage virus (Williams 2003). Since their discovery, many more R-M systems have been identified. The introduction of more efficient techniques for identification has resulted in over 6000 known sequences and over 4000 cloned R-M systems, which include approximately 300 DNA specificities (Vasu & Nagaraja 2013; Mokrishcheva et al. 2011; Roberts et al. 2015).

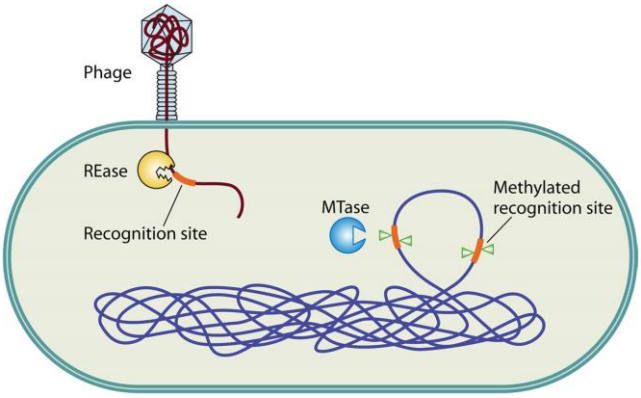


Figure 7: Cartoon diagram of the activity of a generic R-M system (Vasu & Nagaraja 2013).

A typical R-M system includes a restriction endonuclease (REase, also known as a DNA endodeoxyribonuclease or ENase), whose cleavage of DNA is triggered by the recognition of a specific DNA sequence. The other constituent part of the R-M system is an MTase, whose action prevents cleavage of host DNA (Fig. 7) (Bujnicki 2001). In most cases, these functions

are carried out by separate enzymes from the same system (cognate). However, in some systems both of these activities are fulfilled by a multi-subunit protein or even a single peptide (Wilson & Murray 1991). REases and their cognate MTases share the same DNA recognition site, which is normally a specific sequence of around four to eight nucleotides. The modification of the specific site on DNA by the MTase stops it becoming a viable target for its cognate REase. This can also have an inhibitory effect on non-cognate REases. The sequence can be symmetric or asymmetric and can be continuous or interspersed by a defined number of random nucleotides (N) (Wilson & Murray 1991). The genes from which R-M enzymes are expressed often occur at the same loci in the bacterial genome and can be transcribed in sequence or in opposing directions (Williams 2003). It is also possible for enzymes from different systems to share a DNA recognition sequence. The first of these to be discovered is referred to as the "prototype", whilst all other enzymes are known as isoschizomers. Enzymes that share a sequence but cut at different positions within it, are known as neoschizomers. These classifications apply only to REases, and not MTases, due to the difficulty of adequately defining MTase function (Roberts et al. 2003). Generally, R-M systems require a double stranded substrate but there are examples of enzymes that act on single stranded DNA. Most R-M systems are Mg^{2+} dependent and are believed to follow the 2M mechanism (Sträter et al. 1996).

R-M systems are named according to the particular bacterial strain from which they originate. The method follows a three letter code derived from the name of the host organism, taking the first letter of the genus and the first two letters of the species. Systems from the same strain are numbered with reference to the order in which they were discovered (Smith & Nathans 1973). For example, EcoKI was the first of the R-M systems to be identified in *E. coli* K-12 (Murray 2000). The genes that code for these enzymes are most often denoted using this same nomenclature, with an additional subunit determination. For example, the REase gene of EcoKI is *ecoKIR*. The loci of the genes that code for the REase (R) and MTase (M) are often only separated by a few bp and in some cases can overlap. The genes can occur on plasmids, such as with EcoRI, or on viruses, as in the case of EcoPI (Wilson & Murray 1991). At the stage of identification by sequencing, the genes are regarded as encoding putative enzymes, and as such their names are preceded by a "P". This is removed when proof of their activity has been attained (Roberts et al. 2003).

Although similar in enzymatic activity, the R-M systems show great variety in protein structure and gene sequence (Wilson & Murray 1991). To date, there are four classes of R-M

system (Types I to IV), which are separated due to differences in composition, target recognition, cofactors and the manner in which they cleave DNA (Murray 2000). Type I enzymes are large, multi-subunit proteins, which facilitate both restriction and methylation activities. In Type II systems, the two functions are often carried out by separate enzymes. Like Type I systems, Type III enzymes are multi-subunit proteins. They occur as heterotrimers or tetramers, with an M_2R_1 or M_2R_2 stoichiometry respectively. Type III enzymes typically possess a 5 to 6 bp long, asymmetrical DNA recognition sequence, from which they translocate and cleave at the 3' side (Vasu & Nagaraja 2013). The last group is the Type IV systems, which cleave only methylated DNA (Roberts et al. 2003). However, recently discovered were enzymes that cut glycosylated DNA exclusively. As such, the Type IV class encompasses all of the modification-dependent enzymes (Sitaraman 2016). With increasing numbers of and diversity within the known R-M systems, these broad enzyme classes have become less useful. It is also more difficult to assign to them a unique name, as it is possible to have the same three letter code for enzymes from a different genera of bacteria. In an effort to ameliorate this problem, enzymes in one class can be further separated by several subtypes (Roberts et al. 2003). There is also another protein, which is associated with R-M systems. The "C" or "Control" protein is encoded by a gene found upstream from that of the REase, and regulates the expression of the REase in a positive feedback mechanism. In contrast, the MTase gene is expressed constitutively. It is therefore more likely to be able to modify and protect host DNA. Some C proteins can be used to promote the expression of the REases of different bacteria, which suggests the mechanism is derived from a common progenitor (Williams 2003).

Their effect on lambda (λ) phage was the evidence by which R-M systems were first discovered, but also how key characterisation was conducted by Arber and Dussoix in 1962 (Rao et al. 2014). Their work involved inoculating the four *E. coli* strains, K12, B, $15T^-$ and the K12 (P1) lysogen, with λ phage and examining the results. Importantly, they observed that the K12 (P1) lysogen would restrict phage recovered from K12, but not phage used to inoculate the same strain (Arber & Dussoix 1962).With this in mind, it can be reasonably asserted that λ phage was instrumental to the arrival of genetic engineering, as it was the tool used to identify R-M activity (Casjens & Hendrix 2015). Key characteristics of λ made it an ideal model for studying viral infection of bacteria. At 48,502 bp, the bacteriophage is relatively small, such that it was manageable but still comparable to other systems. Like many other viruses, λ can also switch between lytic and lysogenic states (Casjens & Hendrix 2015). These are two different cycles by which a virus uses the cellular machinery of the host to reproduce. The lysogenic cycle involves the integration of the viral DNA into that of the host, a process which

depends on the host survival. In contrast, a lytic virus moves feely in the infected cell and replicates to the extent that it eventually causes cell death. The changeable state of this coliphage has provided a great deal of information about the regulation of viral latency, giving an insight to the activity of human latent viruses, like herpes simplex (Gandon 2016). Phage λ was discovered accidentally in 1951, when K-12 *E. coli* cells were lysed by ultraviolet radiation. It was revealed that most of the laboratory cultures of this strain had been infected by the phage but that this had gone unnoticed. Since then, λ has been the basis for studies on "lysogenicity" (Lederberg & Lederberg 1953). This process is mainly governed by a repressor gene called *cI*, whose protein product activates the expression of its own gene but is able to switch other genes off (Gandon 2016). This maintains the lysogenic state by blocking the transcription of mRNAs necessary for the lytic cycle (Casjens & Hendrix 2015). The lytic cycle is regulated by the product of the *cro* gene. The discovery of other associated proteins (CI and CII) suggests that this system is quite complex, but ostensibly, Cro and CI compete to bind the same operator (Casjens & Hendrix 2015). Although not fully understood, the switch between these two states seems to be determined by the environment in the host cell. Higher numbers of phage make the lysogenic cycle more likely (Gandon 2016). Cyclic AMP (cAMP) knockout mutants showed an increased level of phage lysogeny. As such, cAMP levels are also believed to a determining factor (Hong et al. 1971; Casjens & Hendrix 2015).

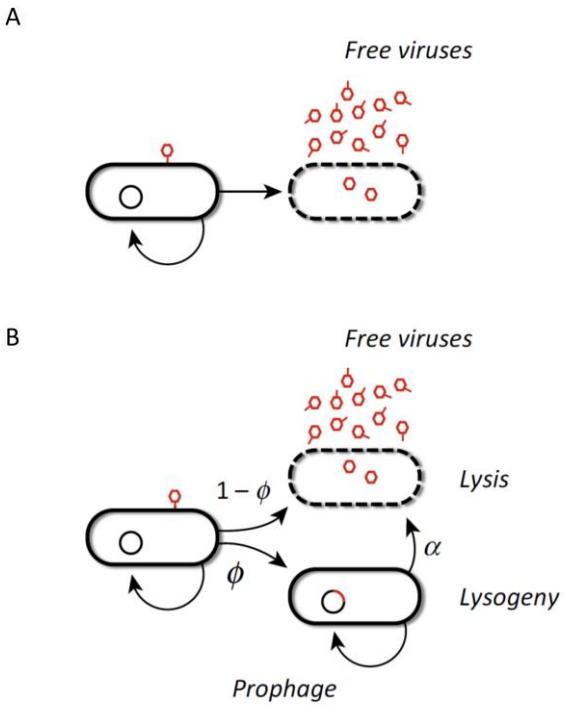


Figure 8: Diagram representing the lytic (A) and lysogenic cycles (B) of bacteriophage λ. Adapted from Gandon 2016.

The sensitivity of phage to restriction can be quantified by "efficiency of plating" (E.O.P.). This value gives the probability of phage survival, by calculating phage plaque formation on a restricting bacterial strain as a proportion of those formed in a non-restricting variant of the same strain. For example, phage infection of a non-restricting bacterial strain would result in an E.O.P. of 1, whereas a restriction active strain can give an E.O.P. of around 10^{-4} (Wilson & Murray 1991). This value varies depending on the R-M system and how many recognition sites are present on the virus. Interestingly, it seems that non-native R-M systems, ie. systems that have been cloned result in a much lower phage E.O.P., often around 10^{-8} (Wilson & Murray 1991). Viruses that successfully avoid restriction are then able to propagate within that same strain. Therefore, subsequent infections will give an E.O.P. of 1. As has been explained, this is due to modification of the virus. With the modification of its own DNA, some viruses will also be modified by the host (Wilson & Murray 1991). Phage also avoid restriction through a lack of recognition sequences, and anti-R-M systems, such as proteins that mimic DNA (McMahon et al. 2009). Smaller viruses are also less likely to contain the recognition sequences of larger REases. Without degradation by REases, phage are able to propagate and therefore duplicate their successful genetic material (Williams 2003).

Given their significant role in protecting the host cell, it is surprising that R-M systems are not essential to prokaryotic life. Knockout mutants are more vulnerable to viral infection but display no other signs of ill-health. As such, R-M systems should be viewed as necessary for the survival of the population, and not the individual cell; R-M activity prevents the spread of foreign DNA to neighbouring cells (Wilson & Murray 1991). R-M systems have other roles, which include involvement in genetic recombination and transposition (Pingoud & Jeltsch 2001). This is of particular significance, as the part played by R-M systems in producing genetic diversity is a subject of increasing concern. More important than their contribution to changes in the host genome are the ways in which the barrier they pose to invading DNA can be overcome. Without this protection, the host genome can be subject to changes caused by Mobile Genetic Elements (MGEs), more specifically by horizontal gene transfer (HGT). This is the phenomenon by which whole genes from outside sources can be incorporated into the genome of an organism, and is proposed as the main cause of the spread of antibiotic resistance (Lindsay 2014). Given this, R-M systems can be regarded as directing evolution. The effect of degrading foreign genetic elements is actually processing them, such that they can be incorporated into the host genome (Bujnicki 2001). Another consideration, is the spread of the systems themselves, and the effect that this can have on bacterial evolution. Sequence homology and codon usage suggests that R-M systems have been transported between

different bacteria by HGT. Their presence on plasmids and viruses is further evidence that they can act as MGEs. The movement of R-M genes can lead to genome rearrangements, the disappearance of specific nucleotide sequences and even host cell death. Together with their primary activity of degrading other foreign elements, R-M systems exhibit behaviour comparable to a "selfish gene". From one perspective, they seem to function as a mechanism for cellular defence, whilst another view is that they are simply insuring their own survival (Kobayashi 2001).

1.5. Type II (Classical) R-M Systems

The defining characteristic of Type II R-M systems, and the most important in terms of their use to molecular biology, is that their REase cleaves double stranded DNA at fixed (easily identified) positions. In most cases this cleavage occurs within their recognition site, but the exceptions cut within 20 bp either side (Marshall & Halford 2010). Some enzymes produce dsDNA with either no overhanging strand (blunt end), or strands protruding by up to 5 nucleotides (sticky ends) (Pingoud et al. 2005). The REase and the MTase in the majority of Type II systems are separate (Marshall & Halford 2010). The MTase is denoted using an "M." prefix, whilst an "R" is added to the name of the REase (although this is often omitted) (Pingoud et al. 2014). On average, the REases are composed of 300 amino acids, whilst MTases are typically 400 (Wilson & Murray 1991). Being relatively small and with few cofactor requirements, these enzymes are compact and efficient molecular machines (Roberts 2005). Despite the thousands of known Type II systems, this only represents a relatively low number of DNA recognition sequences. Given their importance to all walks of biology, there are many efforts to create novel REases with new specificities (Jurėnaitė-urbanavičienė et al. 2016).

Most of the Type II enzymes, such as EcoRI and BglI, conform to the general description of the system as a whole. These "Orthodox" enzymes are ~60 kDa homodimeric complexes, which cleave within or next to their recognition sequence, resulting in a 5'-phosphate and a 3'-hydroxyl end (Pingoud & Jeltsch 2001). As they bind and cleave palindromic sequences, they are known as the Type IIP enzymes (Marshall & Halford 2010). Not all of the Type II enzymes conform to this narrow definition, and so they are separated into other sub-categories (Fig. 9) (Pingoud & Jeltsch 2001; Roberts et al. 2003). The factors that differentiate the many Type II sub-types are: the nature of the recognition sequence, their tertiary and quaternary structure, and the type of cut they produce on DNA (Marshall & Halford 2010; Pingoud & Jeltsch 2001; Roberts et al. 2003).

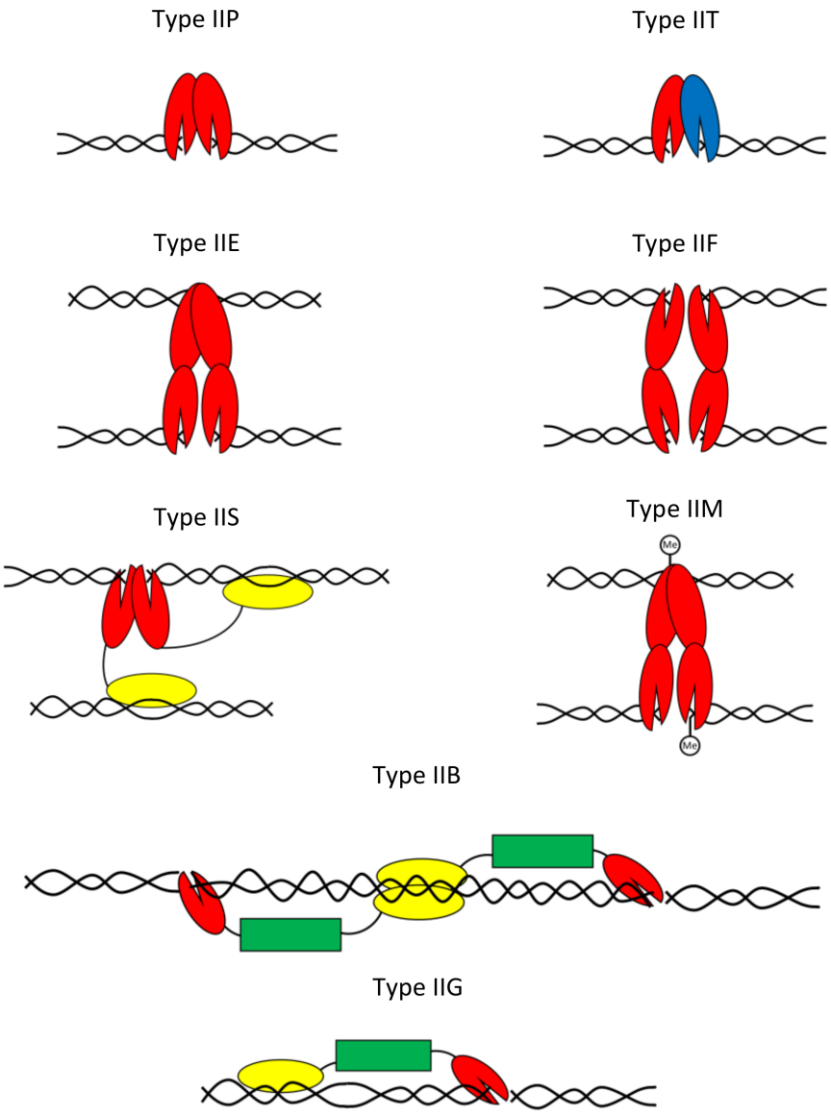


Figure 9: Cartoon showing the subunit organisation and method of DNA cleavage of the Type II REase sub-types. The green boxes represent MTase domains, whilst yellow ovals are TRDs, and "Me" circles on sticks are sites of methylation. Adapted from Bujnicki 2001.

The Type IIS REases have an asymmetric recognition sequence and cleave at set distances away from it (Bujnicki 2001). Characterisation of FokI showed that enzymes in this family consist of two distinct domains and dimerise upon binding DNA (Pingoud et al. 2005). Type IIF enzymes are tetrameric and require two copies of the recognition sequence. They cut at both sites in one action (Bujnicki 2001). The members of the IIE sub-family are homodimers and also need two sites to be active. Interestingly, each monomer has two domains, both of which have a DNA binding site. One of these acts an allosteric site, which binds one copy of the DNA sequence, known as the "effector". The other site binds the "target" sequence and

catalyses DNA hydrolysis. Therefore, this complicated process involves two monomers of the enzyme, each binding two recognition sequences, resulting in cleavage at one site (Bujnicki 2001). Sub-type IIM REases recognise and cleave methylated sequences. A widely known example of this type is DpnI, which has a Type IIE subunit organisation but recognises adenines methylated at the *N*6 position (m6A). It is widely used in the process of site directed mutagenesis to remove non-mutated DNA (Pingoud & Jeltsch 2001; Siwek et al. 2012). The two classes, Types IIB and IIT, both have an unusual subunit organisation. IIT systems are heterodimeric and recognise an asymmetric sequence, although there are exceptions (Pingoud & Jeltsch 2001). Whilst the subunits in IIB systems are also different from one another, they have an organisation and activity not unlike Type I systems. These trimeric enzymes consist of A and B polypeptides, in a A_2B_1 stoichiometry. They possess both their MTase and REase in the one enzyme, and can methylate either symmetrical or asymmetrical sequences. They cleave DNA in a SAM-dependent reaction, and do so either side of their recognition sequence. This results in the removal of a short fragment (Bujnicki 2001; Pingoud & Jeltsch 2001). Like Type IIB systems, IIG REases also possess MTase activity. These enzymes consist of a single polypeptide, which cuts 14 to 16 bp away (depending on the enzyme) from their site in a 3' direction. However, these single chains can methylate only one strand, and so are paired to a cognate MTase, which can modify both (Bujnicki 2001). There is clearly a great deal of variety in structure, topology and sequence specificity within the Type II systems. Nevertheless, they all cleave DNA at a fixed position (Pingoud et al. 2014).

Under normal conditions, the DNA with which a Type II enzyme is interacting would be relatively large. This increases the complexity of the enzyme/DNA interaction. Put simply however, the process begins with non-specific DNA binding by the enzyme, which then randomly diffuses across the DNA. During this step, the enzyme may encounter its recognition sequence, resulting in the activation of phosphodiester bond cleavage. If the enzyme does not come across its recognition sequence, it may dissociate from the DNA. After cleavage has occurred, the enzyme dissociates from the DNA directly or by having transferred across to a non-specific sequence. The dissociation of the REase is the rate limiting step of the reaction, and so will affect its turnover when cutting at multiple sites (Pingoud & Jeltsch 2001). Evidence shows that the sequences either side of the target have an effect on rate. Not only can surrounding nucleotides affect DNA topology, but certain enzymes can accommodate more bases in their active site than just their recognition sequence (Pingoud et al. 2014). DNA-binding proteins that recognise specific sequences can also bind at other positions, but do so with much less affinity. When this occurs, the enzyme undergoes a conformational change,

which involves opening the active site and allowing the DNA to pass through, regardless of whether it contains the recognition sequence. This mechanism has been observed for several DNA-binding proteins, including EcoRV and T7 helicase. When the enzyme is bound to non-specific DNA, there are fewer contacts to the phosphate groups and none to the bases. In general, when bound to their specific sequence, these enzymes reduce in size and bind the DNA more tightly. Of course, the lack of affinity when bound to non-specific DNA enables the enzyme to diffuse along the chain, either by following the contour of a groove in the helix ("sliding") or by moving over the entire chain ("hopping"). It has been proposed that up to 2 X 10^6 base pairs at a rate of 1.7 X 10^6 bp s^{-1} can diffuse through EcoRV upon binding (Williams 2003). Although a change of only one base in the target site results in a significant reduction in binding affinity, cutting at non-specific sites can be observed *in vitro* (Williams 2003). This is known as "Star activity", but only occurs at very high concentrations and under optimum conditions (Pingoud et al. 2014). Other factors which increase its likelihood are the presence of a volume excluder, such as glycerol, or substitution of the divalent metal cofactors. For example, replacing the native Mg^{2+} ions with Mn^{2+} makes binding less specific, but also decreases the rate of the reaction (Williams 2003).

The cognate enzymes in a system recognise and act on the same sequence of DNA, and often, the target base for methylation is next to the point of cleavage (Wilson & Murray 1991). Nevertheless, the enzyme pairs appear to share no sequence homology, even in their target recognition domains (TRDs). It has been proposed that this is due to their difference in quaternary structure. As the MTase is a monomer, the protein must correspond to the complete recognition sequence. Conversely, REases are mostly dimeric, and as such recognise only half the sequence. Their difference suggests that they have evolved independently. MTase function protects host DNA from restriction and prevents cell death. This has selected which REases have survived, and provides a theory for why orphan MTases have no cognate REase (Bickle & Kruger 1993). What is more surprising is that not only is there very little sequence homology between different Type II REases, but that it is rare between isoschizomers. This indicates that their shared specificity is not the product of mutations of a shared precursor, but that these enzymes too have evolved independently. A key example of isoschizomers that are similar, and so are exceptions to this, are EcoRI and RsrI. However, as has been mentioned, EcoRI is encoded on a plasmid. This suggests that this pair share sequence as a result of HGT.

The REases do possess a structural similarity. A five-stranded, mixed β-sheet enclosed by α-helices make up a conserved catalytic core, at the centre of which is a PD…D/EXK motif (Fig.

10). This brings together two carboxylates (one aspartate and one glutamate or aspartate), which are proposed to bind the Mg^{2+} cofactors (Pingoud & Jeltsch 2001). Simple comparisons show that there are often differences in the α-helices and one of the β-strands. The remaining four β-strands are entirely conserved however, two of which hold the PD…D/EXK motif (Pingoud & Jeltsch 2001). The ubiquity of these shared catalytic residues indicates that cleavage by the Type II REases occurs via the same mechanism. Not only can this structure be found in the Type II REases, but also in some other enzymes that possess nuclease activity (Pingoud & Jeltsch 2001). A key example of these is λ-exonuclease, a 5' to 3' single stranded exonuclease encoded by bacteriophage λ (Pingoud & Jeltsch 2001). This enzyme creates 3' sticky ends in the dsDNA of the virus, facilitating recombination and leading to DNA repair (Kovall & Matthews 1998). Despite having an opposing purpose, the conserved structural motif in the λ-exonuclease and the bacterial REases indicates that they are related (Kovall & Matthews 1998). This led to the suggestion that this whole group of REases could be regarded as having developed along one of two branches. Those enzymes that bind the major groove of DNA and produce sticky ends were thought to belong to the "EcoRI family", which includes BamHI and FokI. λ-exonuclease belongs to the "EcoRV family" of enzymes, which move along the DNA minor groove, resulting in blunt-end cleavage. The difference in mechanism is a product of a difference in the topology of the dimer and suggests that this factor is important when considering the evolution of the REase (Pingoud & Jeltsch 2001).

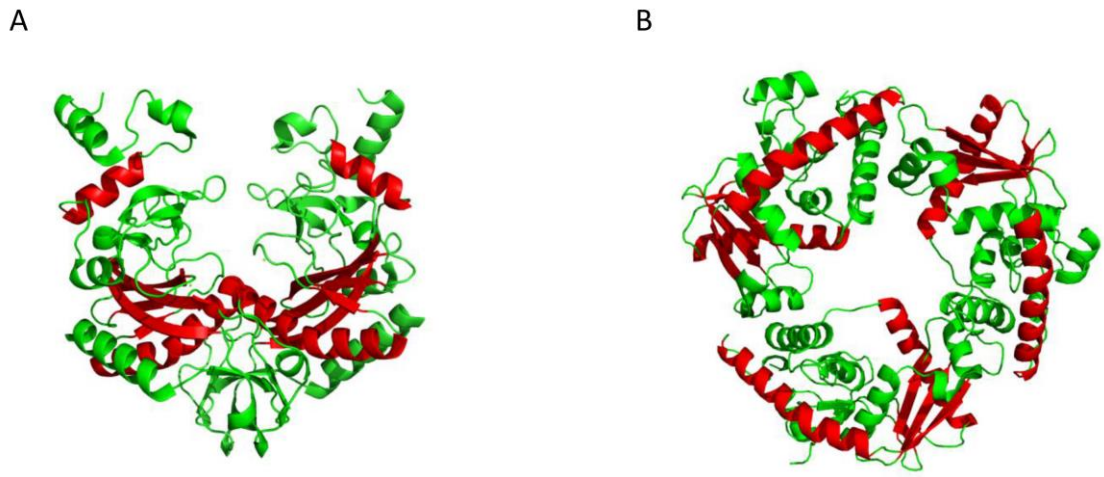


Figure 10: The EcoRV homodimer with two of the conserved catalytic motifs (A) (Kostrewa & Winkler 1995) PDB:1AZ0. λ-exonuclease trimer with three catalytic motifs (B) (Kovall & Matthews 1997) PDB: 1AVQ. The conserved catalytic motifs consist of 2 α-helices and 5 β-strands, and are highlighted in red.

As opposed to the lack of sequence conservation in the REases, there is some within MTases, and in the 5mC MTases in particular (Bickle & Kruger 1993). There are several conserved motifs between large regions of non-homologous sequence. Depending on the limit of definition, there are up 10 motifs within 5mC MTases, 6 of which show higher similarity (Kumar et al. 1994). As would be expected, the amino MTases share a higher degree of similarity with each other than they do with the 5mC MTases (Bickle & Kruger 1993). The TRDs are the most variable region, but similar sequences often produce similar specificities. This is how the TRDs of Type II MTases were first identified (Klimašuaskas et al. 1991).

1.6. Type I R-M Systems

Type I systems were the first R-M systems to be discovered and although not as useful to molecular biology as other types, they are arguably the most interesting (Roberts et al. 2011). These are large hetero-oligomeric complexes, which perform cleavage of DNA away from their recognition site, in an ATP-dependent reaction (Fig. 11) (Murray 2000).

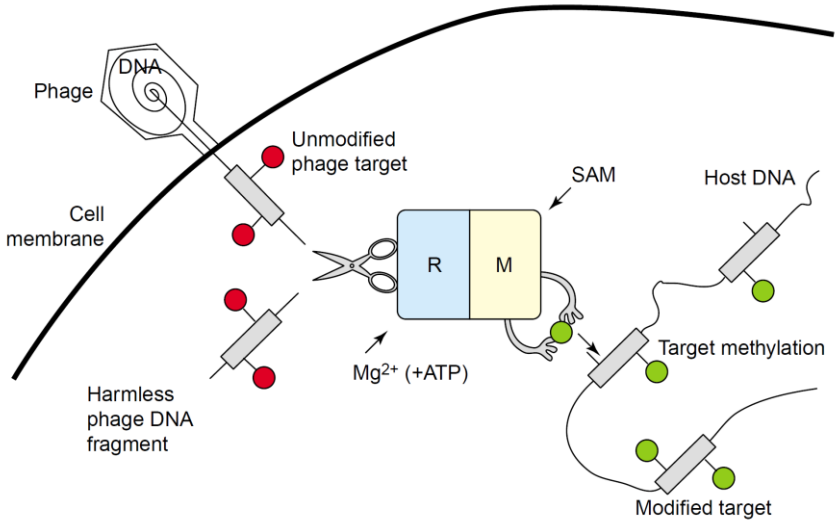


Figure 11: Cartoon diagram of the mode of action of a generic Type I R-M system. Adapted from Tock & Dryden 2005.

A Type I restriction enzyme is composed of a combination of three separate subunits, each having its own role in enzyme activity. These subunits are denoted by Hsd (host specificity for DNA) R, for the restriction subunit, M for the methylase subunit, and S for the specificity subunit (Murray 2000). In general, a Type I system has a ~60 kDa HsdM subunit and a ~130 kDa HsdR subunit, which enables DNA restriction (Taylor et al. 2010). The working ~440 kDa restriction complex has a $R_2M_2S_1$ stoichiometry, whilst a M_2S_1 stoichiometry acts as a cognate MTase for the system (Kennaway et al. 2012). The Type I MTase modifies adenines almost exclusively, and does so on both strands of DNA (Murray 2000). Specific DNA sequences are recognised and subsequently bound by the ~50 kDa HsdS (Taylor et al. 2010). For the most part, these sequences possess the same general organisation of three specific nucleotides followed by a variable spacer of five to eight nucleotides (N), and then a further three to four specific nucleotides (Adamczyk-Poplawska et al. 2011). For example, the target sequence of EcoKI is $AACN_6GTGC$ (Roberts et al. 2011). The S subunit contains two TRDs, separated by a central domain, which is conserved in members of the same family (Adamczyk-

Poplawska et al. 2011). One of the TRDs is specific for one part of the bipartite DNA sequence, while the other TRD recognises the other. The central conserved domain serves to coordinate interactions with the other subunits and more importantly, to separate the TRDs to a defined distance, which corresponds to the non-specific DNA spacer (Powell et al. 2003). This was originally inferred from studying the two Type I systems, EcoR124I and EcoR124II. It is proposed that a recombination event has caused an increase in the conserved sequence and resulted in a change in the nucleotide spacer between the two specific sequences, from 6 bp to 7. This increase of 1 bp was caused by the addition of 4 amino acids to the helical spacer (Murray 2000).

Depending on enzymatic properties, amino acid conservation and the order of the genes from which these systems are expressed, the Type I systems of *E. coli* and *Salmonella enterica* are separated into four sub-families (IA-D) (Roberts et al. 2012). There is relatively little homology between families. For example, EcoKI and EcoAI (from IA and IB families respectively) share only 32% amino acid sequence identity. This indicates that they are indeed evolutionarily linked but that they are further separated than even *E. coli* from *Salmonella* (Murray et al. 1993). The concept of these families originated from the observation that subunits from one Type I enzyme could be used to complement another, and has been a significant step in the understanding of the group as a whole. Unsurprisingly, the biggest region of difference in enzymes from the same family is their S subunit (Murray 2000). Interestingly, it was found that TRDs could be exchanged between common Type I enzymes and doing so would elicit a new DNA specificity. This was a combination of the respective halves of the parental sequences (Janscak & Bickle 1998).

The two TRDs of the Type I MTase do not bind the bipartite DNA target sequentially (i.e. in a 5’ to 3’ direction), with the HsdM subunits positioned above. A circular model has been proposed, which put the TRDs facing each other on opposing strands. The N-terminal and central conserved regions would form a “split linker”, which is interacting with the M subunits (Kneale 1994). TRD 1 binds the top strand of the helix, whilst TRD 2 loops around the DNA via the conserved spacer, and binds the bottom strand in a 3’ to 5’direction (relative to the top strand) (Fig. 12). There were a number of results that led to this conclusion. There is a significant level of homology between the amino acid sequences of N-terminal and spacer domains. In one sub-family, this consists of a repeated sequence, which is split between the two regions (Kneale 1994). This suggested that the N- and C-termini were linked. Truncations of the HsdS that contained only TRD 1 and the spacer, retained their function and bound a

novel DNA sequence. The new target was a palindrome of that recognised by TRD 1, separated by random nucleotides (Abadjieva et al. 1993). This implied the circular arrangement, with the TRDs binding opposite strands. Further evidence for this model was provided by the work of Janscak and Bickle (1998), which created several "permutions" (sic), with covalently fused termini and breaks at different points along the HsdS peptide chain. By creating four structural variants of the HsdS of EcoAI, they showed that engineered C-termini (end points) in the N-terminal and central conserved region had no detrimental effect to DNA binding. These results reinforced the idea that the N-terminal and the spacer were bound to and consequently coordinated in close proximity by the M subunits (Janscak & Bickle 1998).

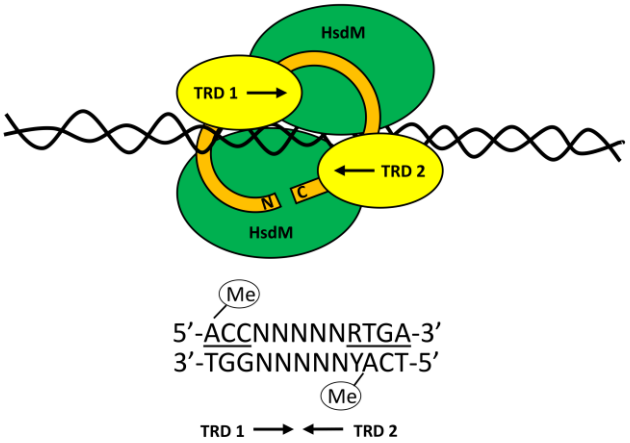


Figure 12: Cartoon model of the subunit organisation of a Type I MTase around its target DNA sequence. Below is an example sequence of the CC398-1 MTase, showing the modified adenine bases on the top and bottom strands (Janscak & Bickle 1998).

Type I enzymes use energy from ATP hydrolysis to translocate DNA. The HsdR subunit binds both the two Mg^{2+} cations and ATP to perform the complicated process involved in producing DSBs in unmethylated DNA (Roberts et al. 2011; Davies et al. 1999). Its N-terminal holds the PD-(D/E)XK motif, and the C-terminal contains several α-helices, which are proposed to facilitate HsdM subunit interaction (Uyen et al. 2009). The HsdR subunit alone possesses no DNA recognition or cleavage properties, although does have a reduced ATPase activity (Šišáková, Weiserová, et al. 2008). For normal activity, it is necessary for two HsdR subunits to bind to the MTase trimer (Holubová et al. 2004). The nuclease domain of HsdR is in a region towards the N-terminal, known as "Region X". Towards the end of Region X, there is also a conserved QXXXY motif. This is also found in RecB nucleases, which are specific to ssDNA. The similarity between the two enzyme classes is difficult to distinguish, and it seems simply that they both translocate DNA and subsequently cleave it non-specifically. It has

therefore been proposed that this shared motif acts as an anchor to the DNA, compensating for the reduced binding affinity for non-specific DNA (Šišáková, Stanley, et al. 2008). Beyond Region X, there are DEAD-box motifs, which are a characteristic of DNA and RNA helicases. These can be found in the HsdR of all known Type I systems. As this is the case, it is widely thought that the translocation of DNA through Type I systems proceeds via a helicase mechanism (Davies et al. 1999). The D-E-A-D box is an ATP binding motif, with the conserved amino acid sequence: (V/I)-L-D-E-A-D-X-(M/L)-L-X-X-G-F (Linder et al. 1989). Mutagenesis studies have shown that these motifs are also essential to the restriction activity of these enzymes (Davies et al. 1999). The HsdR consists of several regions, each important to the different functions of the subunit. Additionally, there seems to be a significant functional relationship between these domains. Mutations at points in the nuclease domain not only eliminate cleavage activity, but also have a negative impact on both ATPase and motor activities (Šišáková et al. 2008). The publication of the crystal structure of the HsdR from EcoR124I, provided a great deal of information and supported previous conclusions about the domain organisation of the subunit. Four distinct domains were highlighted in the solved structure, two helicase (motor) domains above a "helical" (largely α-helices) and an endonuclease domain (Fig. 13). If the helical region interacts with the MTase trimer, and DNA is moved inwards, both the two motor domains and the nuclease domain will face out across the DNA (Lapkouski et al. 2009).

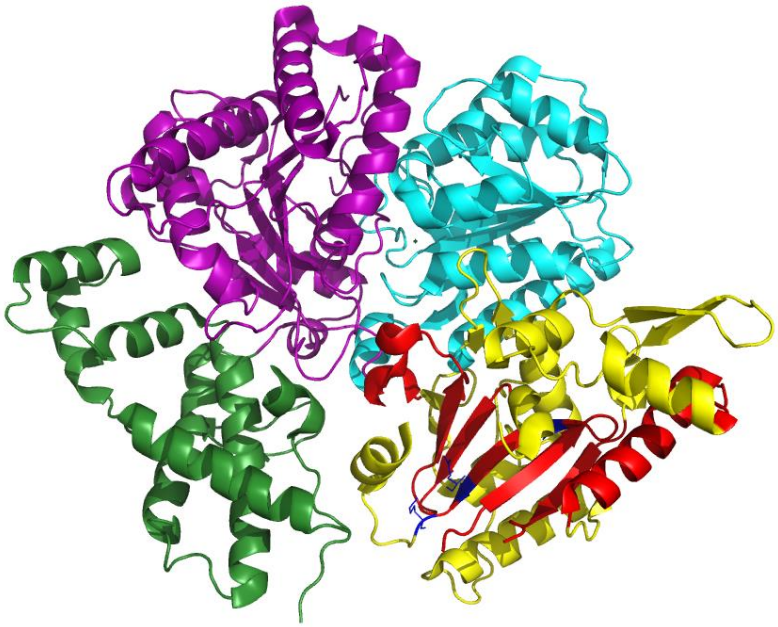


Figure 13: The four domains of the EcoR124I HsdR. The endonuclease domain is highlighted in yellow, the helical region in green and the two motor domains in purple and cyan. The conserved α-helical and β-strand structural motif in coloured red, with the key, Pro-Asp-Lys catalytic motif in blue. EcoR124I HsdR PDB:2W00 (Lapkouski et al. 2009).

The Type I enzyme binds its recognition sequence and the two motor domains reel through the DNA in opposite directions (Davies et al. 1999). The enzyme does not separate the DNA strands, but rather the grooves in the helix act as tracks over which the enzyme moves. The DNA is moved inwards through the enzyme and is thought only to be cut when two HsdR motors collide (Šišáková et al. 2008). This can occur at anything from 40 bp to several kbp away from the recognition site, but is generally about half way between it and the next target site (Davies et al. 1999; Roberts et al. 2011). One molecule of ATP is hydrolysed by the enzyme for every bp translocated (Šišáková et al. 2008). Interestingly, even after DNA cleavage, the enzyme can continue to perform ATP hydrolysis. This can be stopped by cutting the DNA using other enzymes. The protein complex can hydrolyse ATP even when lacking an R subunit ($R_1M_2S_1$). This means that the complex continues to translocate DNA, but is unable to perform cleavage. By observing the translocation process, it was found that the enzymes sometimes became smaller. This, along with the fact that an incomplete complex retains some function, suggested that the R subunits dissociate reversibly from the foundation of a DNA-bound MTase trimer (Roberts et al. 2011). It is this aspect that has led to the conclusion that the Type I systems are able to recycle their subunits, allowing catalytic turnover after cleavage. On linear DNA, it is thought that the R subunits are able to dissociate via the ends and bind a different MTase unit. Conversely, there is no free end on circular DNA and so the HsdR subunits remain bound. In this case, only the MTase can dissociate and can only form a restriction complex when HsdR is in excess (Simons & Szczelkun 2011). Roberts *et al.* carried out further investigations into the conditions required for restriction activity. These elucidated that one or more recognition sequences were required for effective cleavage of circular DNA, whilst at least two sites are needed to restrict linear DNA. However, results from investigations using atomic force microscopy showed that to produce multiple double stranded breaks in DNA that contained multiple sites, the amount of enzyme had to be increased to a 1:1 ratio of enzyme to recognition sequence. As such, Type I systems can be thought of as 'honorary' enzymes, due to their apparent lack of turnover (Roberts et al. 2011).

The consequence of the translocation of up to 50,000 bp, at speeds of up to 1000 bp s^{-1}, is the creation of extruded loops of DNA (Kennaway et al. 2012). These loops are generally very big and can be clearly observed under an electron microscope (Yuan et al. 1980). As DNA is twisted 360^o for every 10 bp of translocation, within these large loops is also a very high degree of positive supercoiling (Yuan et al. 1980; Smith et al. 2009). This type of supercoiling occurs when additional turns to the double helix cause it to wind in on itself (Champoux 2001). This increase in helical density is relaxed when the enzyme eventually cuts the DNA, or if the

enzyme dissociates (Endlich & Linn 1985). Investigations into translocation inhibition concluded that DNA topology does not inhibit the Type I systems, although fluorescence studies have shown that the enzyme can make several attempts before translocation of DNA results in cleavage (McClelland et al. 2005; Janscak et al. 1999).

The three subunits of the Type I system are encoded by the genes, *hsdR hsdM* and *hsdS*, which are expressed from the two promotors, P_{Res} (transcribing *hsd*R) and P_{Mod} (transcribing *hsd*M and *hsd*S). The *hsdM* and *S* genes can be transferred to different bacteria via conjugation or transformation, and they confer non-native modification activity immediately. Equally, the inclusion of the *hsdR* gene will also cause non-native restriction activity. This results in the degradation of the unmodified host DNA, but seems only to be expressed after several generations of the bacteria (Prakash-Cheng & Ryu 1993). Alterations to the TRDs, which are encoded by the *hsdS* gene, can establish a new DNA specificity. Not only can TRDs be swapped, truncated and switched, but also the addition of nucleotides to the region between those that encode them, increases the spacer between the two parts of the recognition sequence (N number) (Murray et al. 1993). Not only do the *hsdM* and *hsdS* genes share the same promoter, but they also share DNA sequence at a junction created by the end of *hsdM* and the start of *hsdS*. Hence, the genes are transcribed from overlapping reading frames and the subsequent translation is coupled. Therefore, during translation, a jump is required to create the two separate polypeptides (Roberts et al. 2012). Roberts *et al.* were able to remove this frameshift from the MTase genes of the Type IA enzyme, EcoKI, to create a fusion of the M and S subunits. Not only did this protein product show full R-M activity *in vivo*, but it could also be over-expressed and purified. With the addition of stoichiometric amounts of EcoKI HsdM protein, the purified fusion formed an active restriction complex *in vitro*. The successful creation of this fusion protein provided a model for an evolutionary intermediate between the Types IA and IB, where the frameshift in the IB systems occurs sooner along the MTase genes (Roberts et al. 2012). The subunits of members of each of the families are approximately the same. This is with the exception of the IB enzymes, which have a smaller M subunit and larger conserved regions either side of their TRDs. This indicates that the evolutionary divergence of the IB systems might have been caused by a gene duplication of a common "half S" ancestor (Fig, 14) (Roberts et al. 2012).

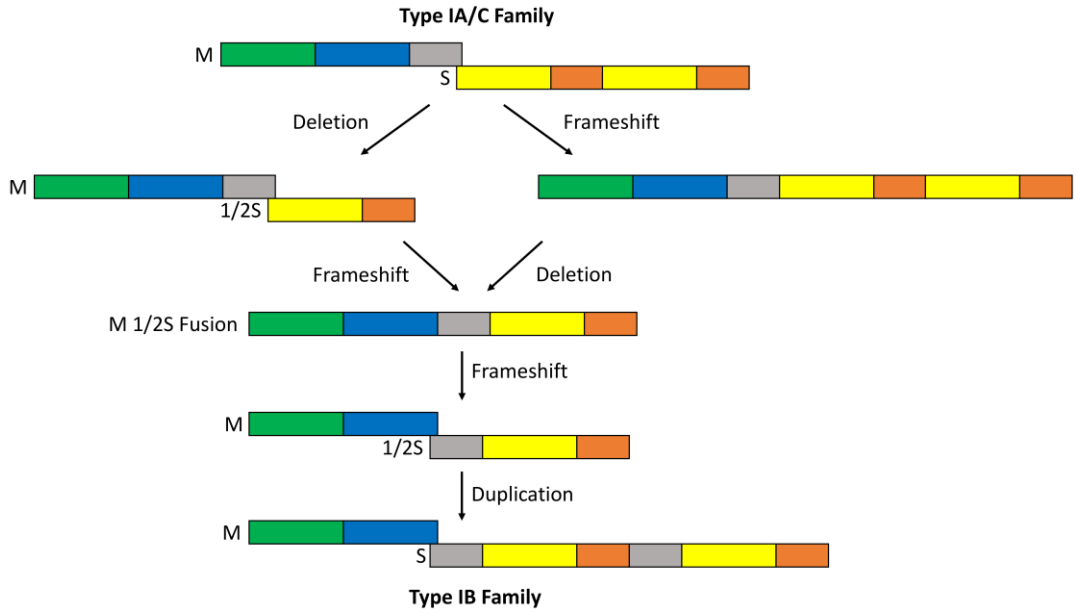


Figure 14: Schematic delineating the changes proposed in the evolution of Type IB systems from Type IA/C systems.
The HsdM is coloured green and blue, where the blue represents the catalytic portion of the subunit. The HsdS is coloured yellow and orange, where the yellow represents the TRDs and the orange blocks are the central conserved and C-terminal sequences. The diagram shows the movement of sequence (grey) from the C-terminal of the Type IA/C HsdM, to the N-terminal of the Type IB HsdS, through subunit rearrangements. Adapted from Roberts et al. 2012.

A relatively recent addition to the Type I sub-families is Type ISP, enzymes from which possess all the functions of a Type I enzyme but within a single polypeptide (SP) (Kulkarni et al. 2016). ISP enzymes were proposed after the discovery of an R-M system from the bacterium *Lactococcus lactis* ssp. *Cremoris* W10 in 2001. Within this organism is a 12.1 kbp plasmid, pEW104, which encodes a single polypeptide with both restriction and modification activities. It was observed that this protein, subsequently known as LlaGI, contained seven DEAD box motifs, and four motifs conserved within adenine methylases. It also possessed a catalytic motif formed by two acidic residues and one basic residue (E-E-K), which can be found in the HsdR of the Type I enzymes, EcoAI and EcoPI (Madsen & Josephsen 2001). Subsequent characterisation of this and its sister prototype, LlaBIII, shows that these systems contain a single TRD, which recognises a 6 to 7 bp asymmetrical target. The enzymes perform ATP-dependent dsDNA translocation, and SAM-dependent modification of single DNA strands. After binding their target, they hydrolyse one to two molecules of ATP for each bp translocated in an exclusively 3’ direction. DSBs are created when target sites are facing each other (5’ to 3’ and 5’ to 3’ on the bottom strand). If only one site is present, or the sites are not in the correct orientation, cleavage activity amounts only to nicking. This was evidence that, like the classical Type I systems, translocation of DNA by the ISPs results in double strand

cleavage only after collision with another motor. In this case, the ISP enzymes are effectively dimerising to initiate restriction activity (Chand et al. 2015). In the N-terminal region of the peptide there is an SF2-like helicase domain, which performs the translocation. This was thought to form large loops of DNA in the fashion of the classical Type I systems (Smith et al. 2009). However, the crystal structure of LlaBIII brought a new perspective on the Type ISP mechanism of translocation. The structure shows an arrangement of six distinct domains, with its MTase at the C-terminal and the TRD at the very end of the protein. This portion of the protein makes the majority of contact to its DNA substrate. It has a structural motif in common with TRD 2 of Type I systems, and the TRD of the Type IIG enzyme, BpuSI. It is responsible for guiding DNA to the upstream ATPase and nuclease domains at the N-terminal of the protein. From this information it was decided that, contrary to belief, the enzyme would not extrude loops of DNA, as movement in this direction would likely displace the nuclease domain. Equally, if the nuclease domain lies upstream from the bound MTase, enzymes translocating towards each other would not have an adequate orientation to produce dimerisation for double strand cleavage. Instead, what is proposed is that the MTase domain feeds the DNA to the helicase, which then allows the MTase and TRD to dissociate from the DNA. As the C-terminal portion is now uncoupled from the substrate, the helicase is able to translocate the DNA. This leads to the movement of the enzyme in a 3' direction, without the creation of DNA loops. Furthermore, this model proposes that DSBs are caused by the combination of nicks by separate ISP peptides (Chand et al. 2015). Interestingly, the crystal structure also includes the target adenine, which is flipped out into the MTase catalytic site (Chand et al. 2015).

As with all the R-M systems, an important aspect of the Type I enzymes is how their defence of the host cell is countered by invading genetic elements. Gene 0.3 of bacteriophage T7 encodes a protein known as OCR (Overcome Classical Restriction) (Fig. 15A). Upon invasion of *E. coli*, OCR is the first protein produced from the genome of the bacteriophage (Stephanou et al. 2009). It mimics the structure of the DNA double helix, and inhibits Type I enzymes by strongly binding to, and blocking their active site. OCR is a negatively-charged homodimer, which mimics a bent DNA structure of about 24 bp long. Normally, DNA is bent by R-M enzymes to allow efficient binding. The structure of OCR is curved by approximately 46^{o}, and is therefore predisposed to bind these enzymes (Walkinshaw et al. 2002). Its negative charge and appropriate shape means that OCR binds with such high affinity it is almost irreversible (50 fold higher than DNA) (Atanasiu et al. 2001). Each of the OCR monomers is 116 amino acids long and has 34 carboxylate amino acids on its surface. In many cases, this combination of Asp and Glu residues directly corresponds to the phosphate backbone of an equivalent DNA

molecule. In order to retain anti-restriction function, each monomer requires at least 16 of these acidic residues (Kanwar et al. 2016). Another example of an anti-restriction DNA mimic is the ArdA (Alleviation of Restriction of DNA A) protein (Fig. 15B). The *ardA* gene is found on plasmids and conjugative transposons in many different prokaryotes (McMahon et al. 2009). An example of this is the *orf18* gene on the Tn*916* conjugative transposon of *Enterococcus faecalis* (Serfiotis-Mitsa et al. 2008; Serfiotis-Mitsa et al. 2010). The crystal structure of ArdA shows that it shares several characteristics with OCR. The dimeric protein has a long, curved, negatively-charged structure, and it mimics a 42 bp long molecule of B-form DNA (McMahon et al. 2009). In contrast to OCR, each ArdA monomer has three α/β domains, across which its charged Asp and Glu residues are distributed (McMahon et al. 2009). The third domain is the dimer interface, and can be severely affected by residue changes (Roberts et al. 2013). Data collected from various ArdA mutants suggest anti-restriction activity is not dependent on dimerisation. ArdA monomers retain the ability to bind the HsdR, and subsequently inhibit DNA cleavage, whereas an ArdA dimer is also able to bind the MTase complex. As such, Domain 3 of this protein is responsible for anti-restriction function, whilst the other two domains coordinate anti-modification activity (G. A. Roberts, Chen, et al. 2013).

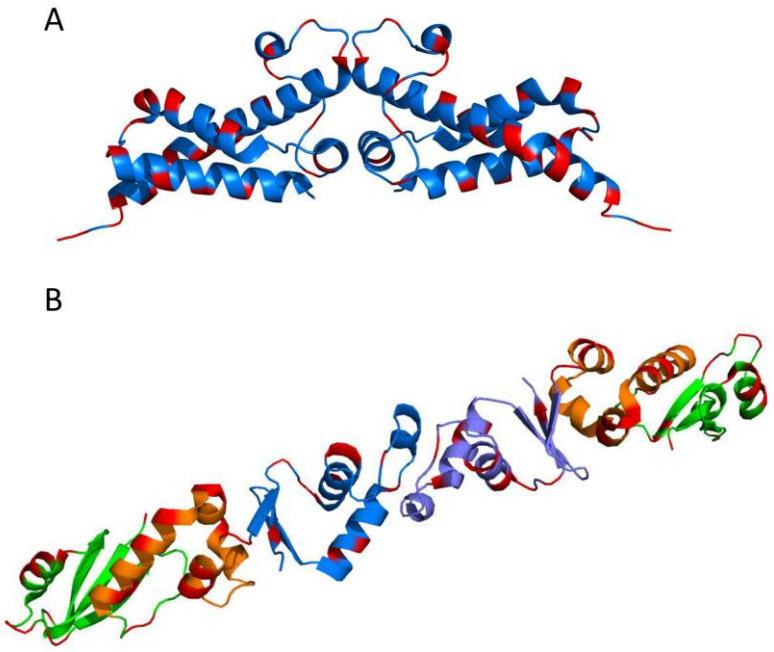


Figure 15: The structure of the OCR dimer (A) PDB: 1S7C (Walkinshaw et al. 2002; Kennaway et al. 2009) and the ArdA dimer (B) PDB: 2W82 (McMahon et al. 2009). For both proteins, the acidic residues are highlighted in red. In the case of ArdA, the two Domain 3s of the dimer are in different shades

1.7. Type I and Type II Systems

Their beneficial characteristics and selfish nature have ensured the survival of the R-M systems. The consequences of intra and inter- genome movements over a significant amount of time have created a wealth of different genes, encoding a multitude of distinct structures. This in turn has meant the evolution of several different types of enzyme, which carry out the same function.

There is clearly a wide structural diversity within the Type II systems. This is partly due to the problems in assigning new enzymes to an already established set of criteria. However, it is undeniable that there are fundamental links between what we call the Type II systems and also similarities to enzymes of different Types. Specifically, there are some Type II structural forms that resemble Type I systems. Enzymes from the unorthodox sub-types, IIB and IIG, share some key characteristics with Type I enzymes. For example, the enzymes from both of these subtypes possess REase and MTase activities in a single enzyme, and are therefore SAM dependent (Bujnicki 2001). Type IIB enzymes, like Type I systems, consist of more than one domain. Type IIB enzymes cut DNA on both sides of their recognition sequence, a defined number of b.p. away. Although Type I systems initially translocate their DNA substrate, they too eventually cut at both directions from their target site (Marshall et al. 2007). In both cases, a small fragment containing the recognition sequence is removed from the rest of the DNA sequence (Marshall et al. 2007). Type IIB systems also have two TRDs in their B, or specificity peptide. This is another direct similarity to a Type I system. In contrast, Type IIG systems possess both activities in a single peptide. They therefore possess only one TRD and cut DNA at one side of their recognition sequence (Bujnicki 2001). For this reason, IIG systems can be viewed as half of the IIB systems. Both sub-types share significant qualities with the Type I systems, but do not translocate DNA. With this in mind, Type IIBs can be considered a "motor-less" type I, and by extension, IIG enzymes can be seen as half of this "motor-less" system (Kennaway et al. 2012).

Kennaway *et al* proposed that the similarities between Types I and II imply that they are evolutionarily linked (Kennaway et al. 2012). Their structural data show that subunit rearrangements of the Type I enzyme, EcoR124I, give it a structure similar to that of a Type IIG enzyme. By removing the motor domains from the EcoR124I HsdR, and joining the C-terminal of the remaining portion to the N-terminal of the HsdM, it has the appearance of the Type IIG enzymes, MmeI and BpuSI. The α-helices that stretch between the two TRDs of the EcoR124I HsdS are absent in the Type II enzymes, which only possess one TRD (Fig. 16).

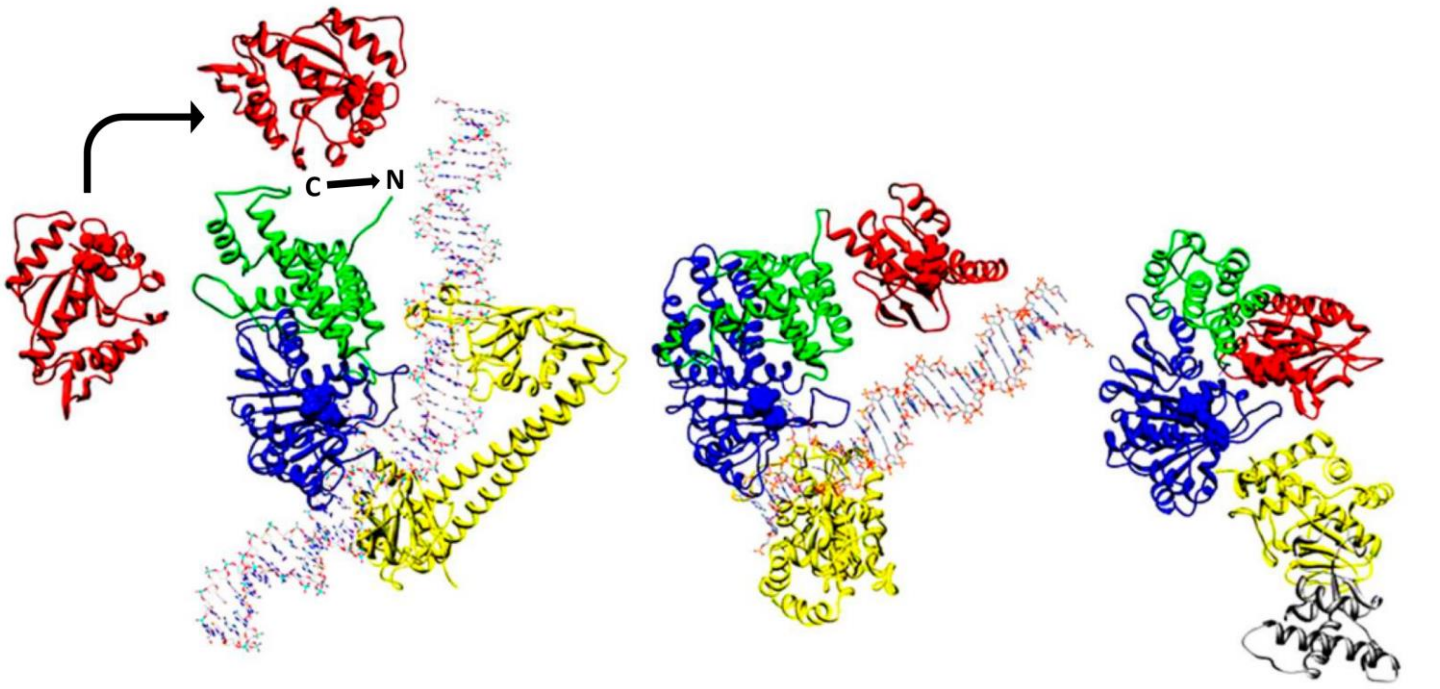


Figure 16: The structure of the Type I enzyme, EcoR124I (far left), with the structures of the Type IIG enzymes, MmeI and BpuSI (middle and right. The EcoR124I structure shows only one of each of its subunits. By joining the C-terminal of the nuclease domain of the EcoR124I HsdR to the to the N-terminal of its MTase, its structure is similar to those of the other two enzymes. All nuclease domains are coloured red, whilst MTase domains are coloured green, blue (catalytic domain), and yellow (TRDs). The BpuSI structure possesses an extra domain (grey), not present in the other two structures. Adapted from Kennaway et al. 2012.

1.8. *Staphylococcus aureus*

Staphylococcus aureus is a gram-positive coccal bacterium, associated with several human diseases (Feil et al. 2003). It is both a prevalent commensal of the skin and nose, and a major pathogen of humans and animals (Kluytmans et al. 1997; Tong et al. 2015). It is estimated that *S. aureus* has colonised approximately 30 % of the human population (Tong et al. 2015). Although mostly carried and not leading to disease symptoms, this bacterium remains one of the major causes of hospital- (HA) and community-acquired (CA) infections and the most common cause of bacteremia (bacterial infection of blood) (Feil et al. 2003; Tong et al. 2015). The precise manner in which *S. aureus* is able to turn from a passive commensal to an aggressive infection is unknown, but it occurs when the immune system of the host is lowered, or there is a breakage of skin or mucosal barrier (Rasmussen et al. 2011). It is therefore known as an opportunistic pathogen (McCarthy et al. 2012).

The fear over increasing case numbers of *S. aureus* bacteremia (SAB) is on account of its severity and its links to secondary infections. In over 30 % of patients, SAB leads to infective endocarditis (Rasmussen et al. 2011). This is an inflammation of heart tissue, particularly the valves, and has a strikingly high mortality rate (Heiro et al. 2006). Another serious consequence of SAB is the spread of infection to the blood brain barrier. This often manifests itself in meningitis, which is an inflammation of the areas around the brain and spine. This also carries a high rate of fatality (Sáez-Llorens & McCracken Jr 2003). SAB most often occurs in immunosuppressed patients, like sufferers of diabetes, HIV, or even patients of an advanced age (Rasmussen et al. 2011). The spread of *S. aureus* and SAB was linked to healthcare environments, predominantly hospitals, where infected individuals are grouped together with non-carriers. However, apparently healthy patients can carry the bacteria and infect the wider community. There are growing numbers of these CA infections. In fact, *S. aureus* is the second most common bacterium found amongst outpatients in the USA (Rasmussen et al. 2011). *S. aureus* infections can also be harmful to animals, which in the case of livestock, frequently carries a financial cost (McCarthy et al. 2011).

First attempts to treat *S. aureus* infections in the 1930s used drugs containing sulphonamide groups. These were soon abandoned due to the increasing bacterial resistance. In the 40s, penicillin became the primary agent used against the bacteria. However, extensive use of the drug brought about the emergence of strains containing a β-lactamase gene, which conferred resistance against penicillin (Fig. 17) (Brumfitt & Hamilton-Miller 1989). By the next decade, the prevailing strains of *S. aureus* were resistant to nearly the entire spectrum of regularly

administered antibiotics. These included erythromycin, streptomycin and all the tetracyclines. With the production of semi-synthetic penicillins, resistant to β-lactamase, it was thought that the worry was over. One of the first of these was methicillin. Unfortunately, it was only a short time after that strains of methicillin-resistant *S. aureus* (MRSA) were identified. Due to the rarity of such occurrences, and that they seemed resistant only to the β-lactam antibiotics, this was not considered an issue. By the 1970s however, MRSA strains resistant to other forms of antibiotics were being detected in Australia. It was at this point that it was realised that MRSA was a significant threat (Brumfitt & Hamilton-Miller 1989). Compared to Europe, there is a higher prevalence of MRSA in the USA (Rasmussen et al. 2011). In the UK and France, techniques such as increased hand-washing and screening, and a rotation of prescribed antibiotics have been able to stem the exponential rise of MRSA numbers (Lindsay 2013; Rasmussen et al. 2011). Nevertheless, MRSA is estimated to cost the European economy 380 million euros each year (Lindsay 2013).

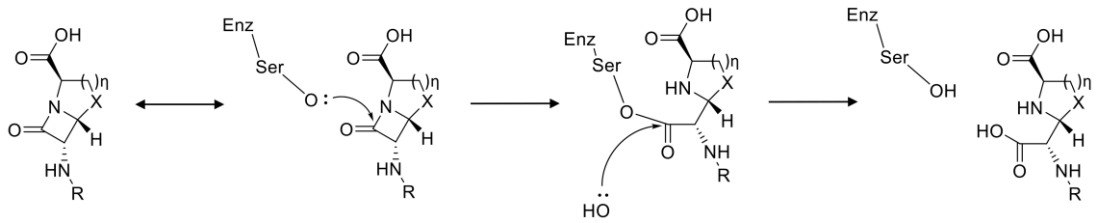


Figure 17: β-Lactam hydrolysis by β-Lactamase. Adapted from Chellat et al. 2016.

More effective DNA sequencing methods enabled the determination of the entire genome of the *S. aureus* Mu50 and N315 strains in 2001 (Lindsay 2014). The genome sequence of 17 other strains, along with the partial sequence of many others has identified that the *S. aureus* genome consists of three groups of genetic material. The first is a highly conserved set of core genes found in all strains. The next division is the MGEs, and makes up approximately 15% of the genome (Waldron & Lindsay 2006). This substantial proportion gives an indication of the regularity of gene transfer between these bacteria. There are also over 700 core variable (CV) genes that make up over 20% of the genome. Due to their uneven distribution, they are the characteristic that defines *S. aureus* lineages (Stefani et al. 2012). Most of the CV genes encode host specific proteins, like virulence factors (Lindsay et al. 2006). Seven of the main CV genes from a large collection of *S. aureus* isolates were sequenced via a process called multilocus sequence typing (MLST). Differences in these genes were used to establish sequence types (STs) for each isolate. Matching profiles between isolates determined them to be clonal, and with at least five of the seven genes shared, they were deemed part of the same

clonal cluster (CC). Amongst the more invasive strains, there were seven genes that encoded proteins for binding human tissue and toxin secretion (Lindsay et al. 2006). This indicated that some strains were more harmful than others (Feil et al. 2003). Another method for strain identification is *spa* typing (Harmsen et al. 2003). The *spa* gene encodes *S. aureus* protein A, a 42 kDa virulence factor, which binds antibodies (Graille et al. 2000). Crucially, this gene contains a repeat region, which is prone to mutation and changes to the repeat number. This variation is used to differentiate *S. aureus* strains, and each is denoted by an alpha-numeric code. Importantly, these do not correlate to CC numbers, but can be used to distinguish variants of the same lineage (Harmsen et al. 2003).

There are ten common *S. aureus* lineages found in humans. These are CC1, CC5, CC8, CC12, CC15, CC22, CC25, CC30, CC45, and CC51 (Lindsay 2010). The predominant clones of HA-MRSA are CC5, CC8, CC22, CC30, CC45, ST239 (Lindsay 2013). Each of these lineages originates from a different part of the World, and has a specific area of circulation. Surprisingly, other than the disparity between the MGEs, isolates from different strains but the same lineage can be genetically very similar. To put this in context, a methicillin-sensitive strain (MSSA) isolated in the UK and an American CA-MRSA strain are both from CC1. Within the core and CV genes (which amounts to around 2500), there were only 285 base differences between the two strains. On the other hand, strains from separate lineages are vastly different (Lindsay 2010). The discrepancy is partly due to homologous recombination, although evidence suggests that this is less likely than point mutations. In contrast to many pathogenic bacteria, such as *Streptococcus pneumoniae* and *Helicobacter pylori*, *S. aureus* is not naturally transformable (Feil et al. 2003). In fact, the only known *S. aureus* capable of accepting *E. coli*-derived plasmid DNA is the chemically engineered RN4220 strain (Veiga & Pinho 2009). However, a huge contributor to lineage diversity is clearly the MGEs, which show signs of high levels of horizontal gene transfer (HGT) and recombination. The main MGEs include bacteriophage and plasmids, which can encode resistance genes and even genes that enable conjugative transfer. A common *S. aureus* MGE is the staphylococcal cassette chromosome (SCC). These are large fragments of DNA, which are exclusively incorporated into the *orfX* gene, and often encode antibiotic resistance (Lindsay 2010). It was discovered that methicillin resistance is conferred by *SCCmec*. This is an MGE that has been successfully incorporated into several *S. aureus* strains, and gives resistance to all β-lactam antibiotics (Lindsay 2013). Interestingly, there are different classes of the *SCCmec* gene, which provide a variety of resistances. The specific gene for methicillin resistance in MRSA is *MecA*. From *MecA* is expressed a protein known as PBP2a, which does not hydrolyse the antibiotic, but

rather inhibits it through competitive binding (Chellat et al. 2016). It is thought that the smaller forms of *SCCmec* are transferred with higher frequency. The range of HGT rates for separate MGEs indicates that some encounter bacterial defence mechanisms (Lindsay 2010).

A notable barrier to HGT between *S. aureus* strains is the R-M systems. The Type II enzymes, Sau3AI and Sau96I have been identified in some lytic strains, and there is even a Type IIB enzyme in the ϕ42 lysogens (staphylokinase-, and enterotoxin A- positive strains) (Veiga & Pinho 2009; van Wamel et al. 2006). The most abundant R-M system amongst *S. aureus* strains is the Sau1 Type I system. This is encoded by one *hsdR* gene and two *hsdM/S* loci on the distant genomic islands, GIα and GIβ (Sung & Lindsay 2007). Two distinct S subunits signifies that the system has two separate DNA target sites (Roberts et al. 2013). Interestingly, a stop codon in the *hsdR* of the transformable RN4220 strain, has rendered it restriction inactive, and therefore transformable. However, complementation with a functional *hsdR* gene restores restriction activity and prevents transformation (Sung & Lindsay 2007). The amino acid sequences of the Sau1 HsdR and HsdM are highly conserved (99%), and so this type of complementation is possible between all lineages (Roberts et al. 2013). The HsdS is not conserved between strains, but is specific to its lineage. Given that Sau1 seems to pose a substantial barrier to HGT, it is suggested that it is more likely to occur within lineages than between them (Waldron & Lindsay 2006). As such, the spread of MGEs is lineage-dependent and each lineage is evolutionarily diverged. This also means that despite the high levels of HGT, the spread of antibiotic resistance is a relatively slow process (Roberts et al. 2013).

Certain MRSA strains are clearly more harmful than others, and antibiotic resistance is variable. Identification of specific infections is therefore extremely important. Due to the hazardous nature of these strains, it is often necessary for analysis to be carried out by experienced researchers, using specialised tools. This however, takes a substantial amount of time, often in situations where it cannot be afforded. There is an undeniable necessity for fast, effective ways of determining *S. aureus* lineages. The Lindsay group has developed a test, which uses PCR to amplify regions of the genome that are unique to the CC. A gene showing high variability is obviously the *hsdS*. By using primers specific to the *hsdS* of each lineage, successful PCR results in strain determination (Stegger et al. 2011).

1.9. *S. aureus* CC398

MRSA from clonal complex 398 (CC398) was first identified in the Netherlands in 2003, and has since been detected with increasing regularity. CC398 is a particular lineage of CA-MRSA, linked to exposure to livestock, chiefly pigs and cattle (van Loo et al. 2007). It is therefore also known as livestock-associated (LA-) MRSA (Ballhausen et al. 2016). Given the extensive transportation of these types of animals, CC398 is likely to have spread considerably, and has been found in Europe, North America and China (van Loo et al. 2007). This is of particular concern after the emergence of strains containing Panton-Valentine leukocidin, a virulence factor which can cause several serious diseases (McCarthy et al. 2011; Lindsay et al. 2006). Consequently, CC398 is regarded as a global health threat (van der Mee-Marquet et al. 2014).

A French study in 2005 that isolated methicillin-sensitive (MSSA) CC398 strains from pig farmers, suggests that CC398 was at first an ordinary commensal of pigs, but extensive use of antibiotics has selected for methicillin resistance. CC398 infection of pigs is seldom dangerous, but can be transferred to humans by simple contact (Armand-Lefevre et al. 2005). A key divergence of CC398 from other *S. aureus* CCs is the acquisition of SaPI-S0385. This is a novel pathogenicity island, derived from the SaPI5 and SaPIbov islands. The new island possesses two putative anti-immune response genes (Ballhausen et al. 2016). Comparison of CC398 isolates from different parts of the World indicates that CC398 actually originated in humans. MSSA CC398 is thought to have transferred to pigs, via the loss of certain MGEs, but was able to develop resistance by acquiring others. It is now able to infect humans as a form of MRSA. This sort of comparison also highlights that there is variation between CC398 isolates from different areas but also different hosts. It was discovered that MRSA CC398 retrieved from humans was more closely related to MRSA CC398 from animal sources, than MSSA CC398 from humans (Lekkerkerk et al. 2015). There is a strong correlation between infection and exposure to pigs, but there is also a significant proportion of infections in people who have had no contact with livestock. Within isolates from these hosts, there is genetic variation (Lekkerkerk et al. 2015). CC398 variants can be identified by *spa* type. The most common are t011, t034 and t108 (Ballhausen et al. 2016). Separate CC398 isolates appear to share the core genes, but as with all forms of *S. aureus*, the majority of difference occurs in the MGEs (Stegger et al. 2010). Some of these MGEs include genes that encode tetracycline and trimethoprim-sulfamethoxazole resistance. It is these genes that can be used to distinguish CC398 variants in the PCR assay.

It is thought that the ability to infect humans is conferred to CC398 by genetic material that has significant homology to bacteriophage φMR-11. The CC398 φMR-11-like prophage is specific to isolates from humans, and is linked to β-converting φ3 prophage (van der Mee-Marquet et al. 2014). This prophage encodes virulence factors, which inhibit human immune response, and enable bacterial transfer from animals to humans. Approximately 90 % of MRSA isolated from humans carries the φ3 prophage. It is thought that CC398 lost this MGE when it transferred to livestock (Ballhausen et al. 2016). This is the basis via which McCarthy *et al.* classified CC398 isolates into two separate groups, which appear to be evolutionarily independent. One group possessed the φ3 prophage, but not the methicillin resistance gene, *mecA*. In the other group, the reverse is true and is thought to have been selected by the widespread use of antibiotics. There is worry over whether the *mecA*-positive strains will acquire the φ3 prophage (McCarthy et al. 2012).

1.10. The Evolution of R-M Systems

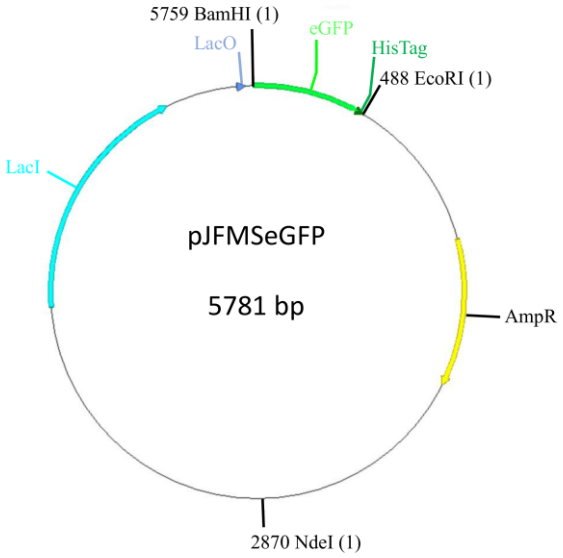
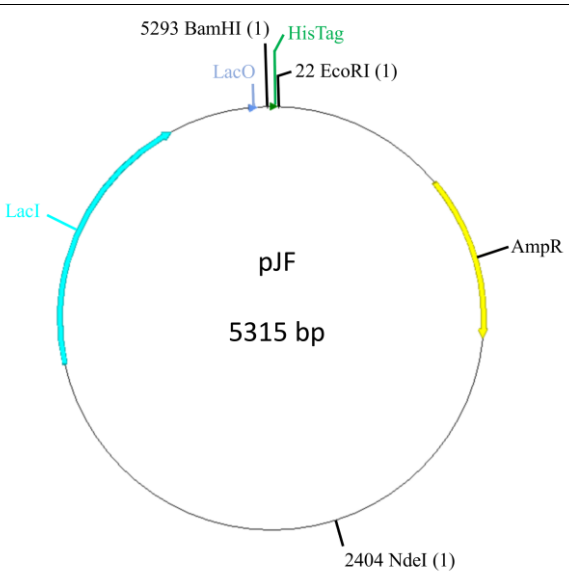
In this thesis, I present work which aims to support the theory that Type I and Type II R-M systems are evolutionarily linked. This proposal was put forward by Kennaway *et al.* in 2012, and is based on the similarities between the Type IIB and IIG subtypes, and Type I systems.

This work gives original findings from the SauI R-M systems and original characterisation of the CC398-1 MTase. Abadjieva *et al.* showed that the second TRD of a Type I HsdS can be deleted, bringing about a change in recognition sequence (Abadjieva et al. 1993). An HsdM to HsdS fusion was created by Roberts *et al.*, using the EcoKI MTase. This was used as the basis for the argument that subunit rearrangements have caused the divergence of the Type IA/C and Type IB subfamilies (Roberts et al. 2012). These findings have not before been used to bolster the argument that Types I and II are evolutionarily linked, and have never before been used in conjunction. Additionally, these investigations were not conducted using a Sau1 system. By making step-wise alterations to the Sau1 Type I subunits from CC398-1 and CC5 systems, I have engineered active enzymes with novel specificities. Soluble, recombinant M to S fusion, half S, and M to half S fusion proteins were successfully produced. These new protein structures are already comparable to Type II enzymes and are only a couple of the proposed subunit rearrangement steps away from them.

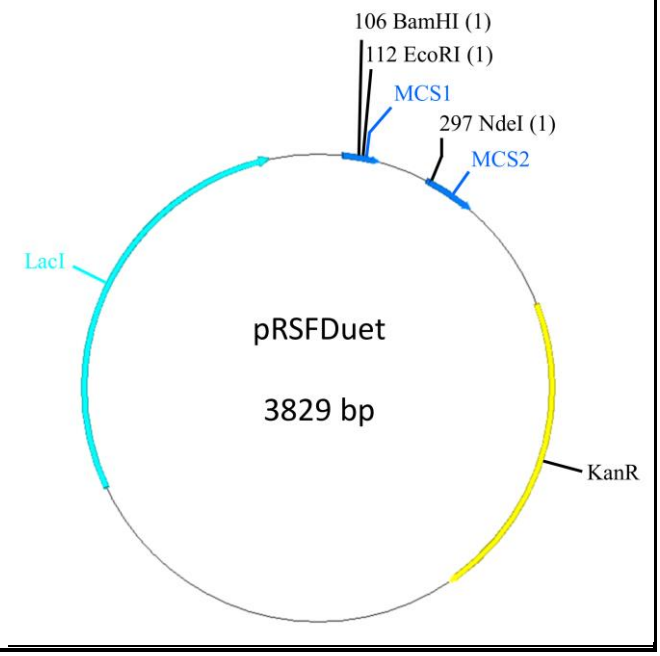
Chapter 2:

Materials and Methods

2.1. Plasmids and Molecular Biology Techniques:

Plasmid	Resistance	Tag	Plasmid Map
pJFMSEGFP	Carbenicillin	GFP and HisTag	
pJF	Carbenicillin	HisTag	

N. B. The pJF vector does not possess a Ribosome Binding Site (RBS). In order to elicit expression, genes to be ligated into pJF must contain the RBS.

Plasmid	Resistance	Tag	Plasmid Map
pRSFDuet-1	Kanamycin	None	

Engineering the pJF vector:

Previous R-M systems had been expressed from a pJFMSEGFP vector, which causes the subsequent recombinant enzyme to carry a HisTag and green fluorescent protein (GFP). For the work shown in this thesis, the GFP gene was removed from the vector by using the polymerase chain reaction (PCR) to amplify all but the GFP sequence of the vector. The following primers were ordered from *Invitrogen* and the Phusion polymerase from *New England Biolabs* (NEB):

pJF Vector Oligonucleotides:

pJFMShis TS- 5'GATCGATCGAGGATCCCATCATCATCATCATCATTAAGAATTC 3'

pJFMSEGFPhis BS- 5'GAGTGAATCCCCGGGGATCCGTCGACC 3'

pJFMShis TS and pJFMSEGFPhis BS primers were used in a PCR containing the following components:

Component	Volume (µL)	Final Concentration (µM)
dH_2O	35.5	N/A
5 X Phusion Buffer (NEB)	10	N/A
dNTP Mix	1	200
pJFMSeGFP Vector Template	1	Variable
pJFMShis TS	1	0.5
pJFMSEGFPhis BS	1	0.5
Phusion Polymerase (NEB)	0.5	1.0 unit

The PCR was carried out under the following conditions:

Temperature (°C)	Duration (seconds)	Cycles
95	300	1
95	30	30
55	30	
72	240	
72	600	1

The PCR product was visualised by agarose gel electrophoresis and excised. A QIAquick Gel Extraction kit (*Qiagen*) was used to elute the DNA, which was then used to transform competent *E. coli* DH5α cells.

Preparing Vector pJF for ligation:

1. Opening the vector

~5 µg of vector pJF was used in a restriction digest with the following components:

Component	Volume (µL)	Final Concentration
dH_2O	76	N/A
10 X reaction buffer	10	N/A
BSA	1	10 µg/µL (1X)
Vector DNA	8	0.625 µg/µL
Restriction enzyme (BamHI, NEB)	5	50 units

The reaction was incubated at 37 ºC for 1 hour.

After the restriction enzyme reaction, a PCR Purification kit (*Qiagen*) was used in order to purify the vector from the rest of the reaction components. It should be assumed that there would be some loss of plasmid material during the purification process and so, given plasmid concentrations are approximate.

The method for PCR purification followed the protocol provided with the Qiagen kit and was as follows:

1. Added “Buffer PB” to the PCR in 5:1 ratio and mixed.
2. Put QIAquick column in 2 mL collection tube.
3. Transferred Sample mix to QIAquick column and centrifuged for 60 seconds at 18320 x *g*. The flow-through was discarded.
4. Added 750 µL “Buffer PE” to QIAquick column and centrifuged for 60 seconds. The flow-through was discarded.
5. Centrifuged sample for another round of 60 seconds.
6. Removed lid of Eppendorf tube with scalpel and placed QIAquick column into the Eppendorf tube.
7. The vector DNA was eluted from the PCR purification column with 50 µL “Buffer EB” and a further 60 seconds in the centrifuge.

2. Calf Intestinal Phosphatase Reaction

As vector pJF had been linearised with only BamHI, the subsequent open ends were complementary to each other and could therefore re-anneal. To prevent this, the phosphate groups at the open ends were removed using Calf Intestinal Phosphatase (CIP).

A "CIP mix" was assembled in the following way:

Component	Volume (µL)
dH_2O	17
10 X "Buffer 3" (NEB)	2
CIP	1 (10 units)

The CIP reaction was then assembled in the following way:

Component	Volume (µL)
dH_2O	51
10 X "Buffer 3" (NEB)	9
Vector DNA	30
"CIP mix"	10 (5 units)

The reaction was incubated at 37 ^{o}C for 1 hour.

After incubation, the reaction mix was treated to PCR purification and eluted with 30 µL of "Buffer EB".

Cloning Techniques:

All primers were ordered from *Sigma-Aldrich* or *Invitrogen* and were used in a PCR reaction using Phusion polymerase (*New England Biolabs*). Colony PCR was conducted using Taq polymerase, unless otherwise stated. All genes were ligated into the engineered pJF vector, using the BamHI restriction enzyme to cut the gene fragment, which was then ligated into the expression vector using T4 DNA ligase (*New England Biolabs*).

Creating Fusion Genes by Crossover PCR

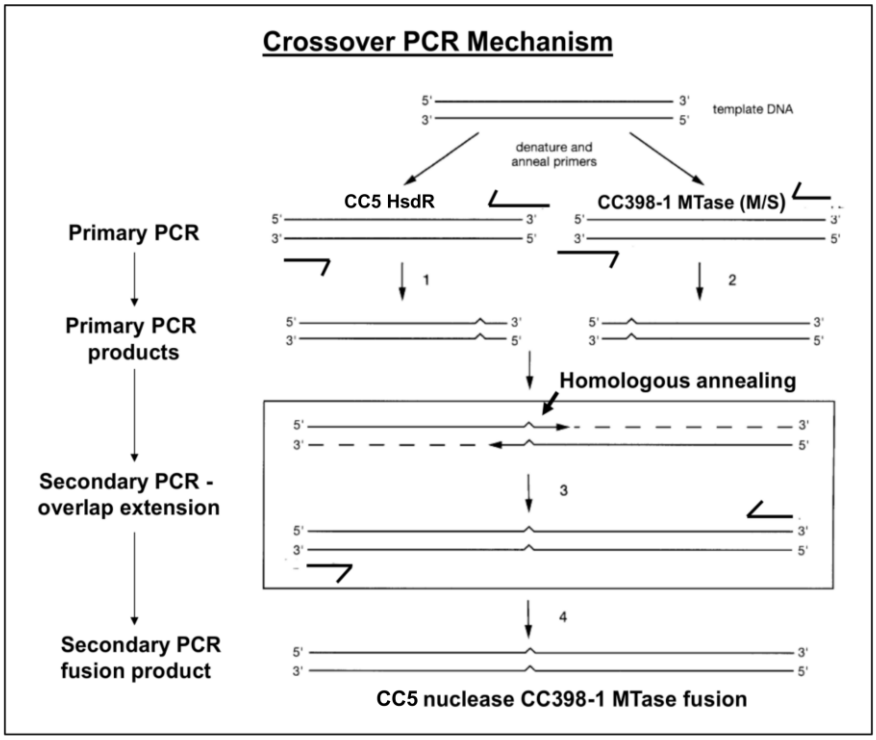


Figure 18: Crossover PCR mechanism

The mechanism of the crossover PCR (Fig. 18) can be summarised by the following:

A primer complementary to the 5'end of the CC5 HsdR ("Primer 1") and another, complementary to both the 3' end of the same gene and the 5' end of the MTase genes ("Primer 2"), were used to amplify a portion of the CC5 HsdR gene, using the CC5 HsdR subunit gene in a pRSFDuet1 vector as a template. In the same round of reactions, a primer complementary to the 3' end of the MTase genes ("Primer 4") and a primer complementary to both the 5' of the MTase and the 3' of the CC5 HsdR gene ("Primer 3"), were used to amplify the CC398-1 MTase genes using the CC398-1 MTase genes in pJFMSeGFP vector as a template. The PCR products from these reactions were used as the templates for a further PCR, using primers 1 and 4. This subsequent reaction creates a product that is a fusion of the two previous products.

Different fusion lengths were created by substituting Primers 2 and 3 in the primary PCR, for primers complementary to different regions of the CC5 HsdR gene.

RM Fusion Oligonucleotides:

All *hsdR* to *hsdM* gene fusions were created using the *hsdR* from the N315 (CC5) strain of *S. aureus,* and the *hsdM* from a CC398 strain of *S. aureus,* as templates. Therefore, the forward primer for the endonuclease primary PCR and the reverse primer for the specificity subunit (MTase) primary PCR, were the same for all of the constructs.

The nucleotide sequences were as follows:

Mu50nucuni TS (Primer 1)

5' AGTCAGTCAGGGATCCAAGAAGGAGATATACAT**ATG**GCATACCAAAGTGAATACGC 3'
BamHI RBS Start **Endonuclease**

CC398-1BS (Primer 4)

5' GATCGAATTCC**GGATCC**AATAAACATCTTTTGAAGTAATGAC 3'
BamHI **Specificity subunit**

Fusion Name	Label	Nucleotide Sequence
RM_CM_1*	Mu50nucshort-CC398-1ts	5'**GAAACAGATAGAATACTGATGGC A**ATGTCTATTACTGAAAAACAA 3'
	Mu50nucshort-CC398-1bs	5'CGTTGTTTTTCAGTAATAGACAT**TG CCATCAGTATTCTATCTGTTTC** 3'
RM_CM_2*	Mu50nuclong-CC398-1ts	5'**CCGTATCAAGTGTATGCGGTAGA A**ATGTCTATTACTGAAAAACAA 3'
	Mu50nuclong-CC398-1bs	5'CGTTGTTTTTCAGTAATAGACAT**TT CTACCGCATACACTTGATACGG** 3'
RM_EB_1	Mu50nucendalpha-CC398-1TS-	5'**GCACTTATTCAACAAGCGACT**AT GTCTATTACTGAAAAACAACG 3'
	Mu50nucendalpha-CC398-1BS-	5'CGTTGTTTTTCAGTAATAGACAT**AG TCGCTTGTTGAATAAGTGC** 3'
RM_EB_2	Mu50nucendcoil-CC398-1TS	5'**CGACTGAGACAGGGAATAAT**ATG TCTATTACTGAAAAACAACG 3'
	Mu50nucendalpha-CC398-1BS	5'CGTTGTTTTTCAGTAATAGACAT**AT TATTCCCTGTCTCAGTCG** 3'
RM_EB_3	Mu50nuclease-CC398-1TS	5'**CATTTGCTGAGTCGTTTATGAGA** ATGTCTATTACTGAAAAACAACG 3'
	Mu50nuclease-CC398-1BS	5'CGTTGTTTTTCAGTAATAGACAT**TC TCATAAACGACTCAGCAAATG** 3'

Fusion Name	Label	Nucleotide Sequence
RM_EB_4^	RM_EB_7 TS	5'**CAGAATAACCGAATCAATACA**ATG TCTATTACTGAAAAACAACG 3'
	RM_EB_7 BS	5'CGTTGTTTTTCAGTAATAGACAT**TGT ATTGATTCGGTTATTCTG** 3'
RM_EB_5^	RM_EB_8 TS	5'**CAATCATTTGCTGAGTCGTTT**ATG TCTATTACTGAAAAACAACG 3'
	RM_EB_8 BS	5'CGTTGTTTTTCAGTAATAGACAT**AAA CGACTCAGCAAATGATTG** 3'
RM_EB_6^	RM_EB_9 TS	5'**CCCTGTCAATTAGCTAAGATGATA** ATGTCTATTACTGAAAAACAACG 3'
	RM_EB_9 BS	5'CGTTGTTTTTCAGTAATAGACAT**TAT CATCTTAGCTAATTGACAGGG** 3'
RM_EB_7^	RM_EB_10 TS	5'**CTGATGGCAATGCGTCCGTAT**ATG TCTATTACTGAAAAACAACG 3'
	RM_EB_10 BS	5'CGTTGTTTTTCAGTAATAGACAT**ATA CGGACGCATTGCCATCAG** 3'
RM_EB_8^	RM_EB_11 TS	5'**CCGCTACATACAAATGTTT**ATGTCT ATTACTGAAAAACAACG 3'
	RM_EB_11 BS	5'CGTTGTTTTTCAGTAATAGACAT**AAA CATTTGTATGTAGCGG** 3'
RM_EB_9^	RM_EB_12 TS	5'**GGTGTTGAAACGCGATACTTTTCT** ATGTCTATTACTGAAAAACAACG 3'
	RM_EB_12 BS	5'CGTTGTTTTTCAGTAATAGACAT**AG AAAAGTATCGCGTTTCAACACC** 3'
RM_EB_10^	RM_EB_13 TS	5'**GGCATACAACTGGAAGTGGT**ATGT CTATTACTGAAAAACAACG 3'
	RM_EB_13 BS	5'CGTTGTTTTTCAGTAATAGACAT**ACC ACTTCCAGTTGTATGCC** 3'
RM_EB_11^	RM_EB_14 TS	5'**GCGAGTCAGATTTTATCA**ATGTCT ATTACTGAAAAACAACG 3'
	RM_EB_14 BS	5'CGTTGTTTTTCAGTAATAGACAT**TGA TAAAATCTGACTCGC** 3'

* Gene fusion created by Christopher McLean

^ Gene fusion created by Dr John White

MTase Oligonucleotides

Changes to the CC398-1 MTase genes were made using PCR, with the CC398-1 MTase gene as a template and the following oligonucleotides:

Construct Name	Label	Nucleotide Sequence
CC398 HsdM	HsdM-TS	5'GATCGATCGGATCCAAGAAGGAGA TATACATATGTC 3'
	HsdM-BS	5'CGATCGGATCCTTACTCATCTTTCAA CACCCCAAGTTC 3'
MS fusion	HsdM-TS	5'GATCGATCGGATCCAAGAAGGAGA TATACATATGTC 3'
	MTasefusion-TS	5'CTTGGGGTGTTGAAAGATGAGATGA GTAATACACAAAAGAAAAATGTGC 3'
	MTasefusion-BS	5'GCACATTTTTCTTTTGTGTATTACTC ATCTCATCTTTCAACACCCCAAG 3'
	CC398-1BS	5'GATCGAATTCCGGATCCAATAAACA TCTTTTGAAGTAATGAC 3'
HalfSHis	HsdM-TS	5'GATCGATCGGATCCAAGAAGGAGA TATACATATGTC 3'
	CC398-1 TRD 1 BS	5'GATCGAATTCCGGATCCCAATTCTT GCGAGAAGATTTTCTGC 3'
CCHalfSHis	HsdM-TS	5'GATCGATCGGATCCAAGAAGGAGA TATACATATGTC 3'
	CC398 -1 TRD 1 BS2	5'GATCGAATTCCGGATCCATCTTTAC CATTCTCATCTTTAAATCG 3'
HalfS	HsdM-TS	5'GATCGATCGGATCCAAGAAGGAGA TATACATATGTC 3'
	CC398-1 TRD 1 BS wSTOP	5'AATTCCGGATCCTTACAATTCTTGC GAGAAGATTTTC 3'
CCHalfS	HsdM-TS	5'GATCGATCGGATCCAAGAAGGAGA TATACATATGTC 3'
	CC398-1 TRD 1 BS2 wSTOP	5'AATTCCGGATCCTTAATCTTTACCAT TCTCATC 3'
MCCHalfS fusion[¥]	HsdM-TS	5'GATCGATCGGATCCAAGAAGGAGA TATACATATGTC 3'
	CC398 -1 TRD 1 BS2	5'GATCGAATTCCGGATCCATCTTTAC CATTCTCATCTTTAAATCG 3'

*The MS fusion was created using the "cross-over" PCR method.

[¥] The M to CCHalfS fusion was created using MS fusion as a template.

Ligation of DNA

Ligation reactions were mostly conducted by taking an aliquot of open sample vector (of variable volume, depending on concentration) and adding insert DNA. T4 DNA Ligase (*New England Biolabs*), Ligase buffer and nuclease-free water. The ligation mixtures were incubated at room temperature for ~10 minutes. The mixture was then directly used for transformation of competent *E. coli* DH5-α cells.

Ligation reactions were generally carried out using the following protocol:

Reagent	Sample Volume (µL)	Concentration
Sample vector	3	0.1 µg/µL
Sample insert	7	Variable (3-4 X vector concentration)
Ligase buffer	2	N/A
Nuclease-free water	7	N/A
T4 DNA ligase (NEB)	1	400 units

Volume of insert could vary and total volume of the reaction was adjusted by an appropriate increase or decrease of the water component.

Agarose Gel Electrophoresis of DNA

Agarose gel electrophoresis was used to visualise and identify specific lengths of DNA. Agarose gels (0.8-1.0% (w/v) were cast (as horizontal slab gels) and run in 1x TAE buffer (40 mM Tris·Cl, 2 mM EDTA, 24 mM acetic acid, pH 7.7) as described by Sambrook & Russell, 2001. Ethidium bromide was included in the molten agarose at a final concentration of 0.5 µg/ mL. Samples were prepared in 1X DNA loading buffer (25% v/v ficoll 400, 0.25% v/v xylene cyanol, 0.25% v/v bromophenol blue) and loaded into wells using a Gilson pipette. Samples were run alongside molecular weight markers at 100 V/h. DNA was visualised using a UV transilluminator, and images were acquired using a digital camera.

Recovery of DNA from Agarose Gels

DNA was extracted from agarose gels by using a QIAquick Gel Extraction Kit (*Qiagen*), a microcentrifuge and following the protocol provided by the manufacturer. The DNA was eluted from the QIAquick purification columns in either 30 or 50 µl of "Buffer EB".

Colony PCR

Colony PCR was used to confirm the correct orientation of genes ligated into the pJF expression vector. Four colonies were picked from an LB agar plate and suspended in 20 µL of dH_2O. 10 µL of each sample was used to inoculate 5 mL of LB broth (used to establish an overnight culture), whilst the remaining 10 µL was used in a PCR containing:

Component	Volume (µL)	Final Concentration (µM)
Colony Sample	N/A	N/A
pJFMS Promoter Forward Primer	1	1
Reverse Primer	1	1
10 X Taq Buffer	5	N/A
$MgCl_2$	3	1500
dNTP Mix	1	200 (each dNTP)
1 X Taq Bead (*GE Healthcare*)	N/A	N/A
dH_2O	39	N/A

The PCR was carried out under the following conditions:

Temperature (°C)	Duration (seconds)	Cycles
95	300	1
95	30	30
55	30	
72	240	
72	600	1

Preparation of Plasmid DNA

A QIAprep Spin Miniprep Kit and a microcentrifuge were used. Cells from 5-10 mL overnight culture of *E. coli* DH5α (that had been previously transformed with the construct of interest) in LB medium (supplemented with 100 µg/ mL carbenicillin), was pelleted by centrifugation in a 15 mL Falcon tube at 1520 x *g* for 10 min. A Miniprep (*Qiagen*) kit was used to extract and purify the plasmid DNA from the cells.

The protocol contained in this kit was carried out without alteration, and was as follows:

1. Bacterial cell pellet was resuspended in 250 µL "Buffer P1" and transferred to an Eppendorf tube.
2. To this mix was added 250 µL "Buffer P2". This was inverted 6 times to mix.
3. To this mix was added 350 µL "Buffer N3". This was inverted 6 times to mix.
4. The sample was centrifuged for 10 minutes at 18320 x *g*.
5. The supernatant was transferred to a QIAprep spin column and centrifuged for 60 seconds. Flow-through was discarded.
6. 500 µL "Buffer PB" was added to the column, which was then centrifuged for another 60 seconds. Flow-through was discarded.
7. 750 µL "Buffer PE" was added to the column, which was again centrifuged for 60 seconds. Flow-through was discarded.
8. The column was centrifuged for a further 60 seconds. Flow-through was discarded.
9. Lid of Eppendorf tube was removed with a scalpel and placed QIAprep spin column into the Eppendorf tube.
10. 50 µL "Buffer EB" was added to the QIAprep spin column and left to stand for 60 seconds. The plasmid DNA was then eluted by a further 60 seconds in the centrifuge.

The ~50 µl of plasmid solution (~50 ng/µl) that resulted from this process was stored at either 4°C or -20°C.

Diagnostic Restriction digestion

Ligation reactions were screened by diagnostic restriction BamHI (*New England Biolabs*) digests. These reactions were carried out using the following protocol:

Reagent	Volume (µL)
10X reaction buffer	1
100X BSA	0.5 (10X)
Sample plasmid	8
Restriction enzyme	0.5 (5 units)

The results of these digests were analysed by agarose gel electrophoresis and visualised by UV light. A positive result would show a band of high molecular weight, corresponding to the plasmid vector (~5315 bp, pJF) and a smaller fragment (2754 bp, CC398 MTase), corresponding to the insert.

Phage lambda preparation

Fresh dilutions of lambda (λ) phage were prepared using a Dryden lab λ phage stock. Comparing previous spot tests led to an estimate of the appropriate phage titre (plaque-forming units/ mL) to use for inoculations. A single colony of the *E. coli* NM1261 strain was used to inoculate 5 mL of LB (supplemented with 0.2% Maltose, 10 mM $MgSO_4$, 100 µg/ mL carbenicillin) in a 50 mL Falcon tube, and the culture was grown at 37 °C overnight, whilst shaking. 100 µL of λ phage (of the appropriate phage titre) was added to 200 µL of the *E. coli* overnight culture and incubated at 37 °C for 15 minutes. 3 mL of molten BBL top agar (maintained at 42 °C) was added to the cell mixture, which was then poured and spread evenly onto BBL bottom agar plates (supplemented with 100 µg/ mL carbenicillin). The plates were incubated at 37°C overnight. After the overnight incubation, a 5 mL pipette was used to flood the plate with 5 mL of phage buffer, which was left to soak for half an hour at room temperature (25 °C). After this incubation, it was assumed that the phage had been resuspended in the phage buffer, which was then carefully removed with a pipette and centrifuged at 1920 x *g* for 10 minutes. A pipette was used to remove the top 4 mL of buffer, to which 50 µL of chloroform was added and mixed, in order to kill any residual bacterial cells. This was stored at 4 °C in a sealed, labelled container. Spot tests were used to identify the relative concentration of different phage stocks and the stocks were adjusted with additional phage buffer accordingly.

2.2. Gene sequencing and SMRT

1 µL of primer (25 µM) was added to 5 µL of sample plasmid for each reaction. Sequence reads were reliable to at least 600 bp. All big dye reactions and subsequent gene sequencing was performed by *Genepool* (The University of Edinburgh), the results of which were analysed using FinchTV (FinchTV 2016) and ApE software.

Bacterial Genome Sequencing

Competent *E. coli* ER2791 cells were transformed with a plasmid containing the target MTase, and spread on a plate of LB agar (supplemented with 100 µg/mL carbenicillin). Plates were incubated at 37 ^{o}C, upside down, overnight. Genomic DNA was harvested and purified from the ER2791 cells in either of two methods, using the Wizard ® Genomic DNA Purification Kit (*Promega*) or Phenol/Chloroform extraction.

This process was started by picking a successfully transformed colony into 5 mL of LB (supplemented with 100 µg/ mL carbenicillin), and incubating overnight at 37 ^{o}C, whilst shaking. Cells from the subsequent culture were separated into 1 mL aliquots and harvested by centrifugation at 2380 x *g* for 15 minutes. The Wizard ® Genomic DNA Purification Kit (*Promega*) was then used to lyse the cells and purify the genomic DNA.

The protocol contained within the Wizard ® Genomic DNA Purification Kit (*Promega*) was carried out without significant alteration. The protocol describes a method for gram negative and gram positive bacteria. competent *E. coli* ER2791 cells are gram negative, and so the method was as follows:

1. 1 mL cell pellet was resuspended in 600 µL of "Nucleic Lysis Solution".
2. Sample was the incubated in a water bath at 80^{o}C for 5 minutes.
3. Sample left on bench top to cool at room temperature.
4. 3 µL of "RNase Solution" was added to sample, which was then inverted 6 times to mix.
5. Mix incubated at 37^{o}C for 60 minutes.
6. Sample left on bench top to cool at room temperature.
7. 200 µL "Protein Precipitation Solution" was added to sample, which was the vortexed carefully but at high speed. Protocol suggests vortexing for 20 seconds, longer was required, but was conducted with caution.
8. Sample left on ice for 5 minutes.

9. Sample was centrifuged at 18320 x *g* for 5 minutes.
10. Cell debris and protein created a pellet, but this did not adhere strongly to Eppendorf wall. Supernatant was carefully transferred to clean Eppendorf, containing 600 µL room temperature isopropanol.
11. Sample was inverted several times to mix. Genomic DNA became visible.
12. Sample was centrifuged at 18320 x *g* for 4 minutes.
13. DNA created small white pellet. Liquid supernatant was carefully discarded.
14. 600 µL 70 % ethanol added to Eppendorf, to wash pellet. Tube inverted several times carefully.
15. Sample was centrifuged at 18320 x *g* for 2 minutes.
16. Liquid supernatant was carefully discarded and pellet left to dry for 15 minutes with Eppendorf open, at room temperature on benchtop.
17. 100 µL "DNA Rehydration Solution" added to pellet, and sample left overnight at 4ºC.

Phenol/Chloroform extraction was carried out using the following protocol:

A successfully transformed colony was picked from an agar plate and used to inoculate 15 mL of LB (supplemented with 100 µg/ mL carbenicillin), which was then incubated overnight at 37 ºC, whilst shaking. The subsequent saturated cell culture was centrifuged at 10967 x *g* for 5 minutes at 4 ºC. Whilst on ice, the supernatant buffer was separated from the cell pellet, which was then resuspended in 1100 µL Resuspension buffer (25% sucrose, 50 mM Tris pH 8.0, 1 mM EDTA). 200 µL of fresh lysozyme (supplemented with 10 mg/ mL in 0.25M Tris pH 8.0) was added, mixed and then left on ice for 1 to 2 hours.

After incubation on ice, the lysed cell culture was checked for viscosity. 800 µL of lysis buffer (1% TritonX-100, 50 mM Tris pH 8.0, 62 mM EDTA) and 200 µL 10% SDS was added and gently mixed with the cells, to complete the lysis.

The lysed cell mix was the poured into a 50 mL centrifuge tube and 1 volume of phenol (2.5 mL): chloroform (2.5 mL): isoamyl alcohol was added. This mix was then shaken until homogenous and centrifuged at 17136 x *g* for 10 minutes. The centrifuged sample had separated into different phases. 800 µL of the aqueous phase was drawn with a pipette and ejected into a 1.5 mL Eppendorf tube, to which was added an equal volume of phenol. This was then centrifuged at 17136 x *g* at 4 ºC for 10 minutes. The aqueous phase was transferred

into a new Eppendorf, an equal volume of methylene chloride was added and then this was centrifuged at 17136 x *g* at 4 ^{o}C for 10 minutes. This step was repeated. The aqueous phase of this final step was transferred into a new Eppendorf tube, to which was added 1: 10 volume 5 M NaCl and 0.7 volume isopropanol. A clean capillary tube was then used to hook the now visible fibres of DNA and twist them around the tube. The DNA could be taken out of the solution and washed by gently putting the capillary end into an Eppendorf containing cold 70% ethanol. This was repeated with fresh 70% ethanol. The DNA was then left to dry on the capillary tube for 5 minutes and then dissolved in 400 μL of "Buffer TE" (*Qiagen*).

Pacific Biosciences SMRT Sequencing

Preparation of genomic DNA for SMRT sequencing by *Pacific Biosciences* was carried out in the laboratory of Dr. Richard Morgan (*New England Biolabs*). All consumables, including those specific to *Pacific Biosciences* SMRT sequencing, were supplied courtesy of Dr. Richard Morgan and *New England Biolabs*.

The protocol for this procedure is specific to each experiment and was altered under expert guidance by Dr. Richard Morgan and Yvette Luyten (*New England Biolabs*).

2.3. Competent cells and Gene expression

Cell lines used:

Cell Line	Cell Genotype	Supplier
BL21	F^- ompT gal dcm lon $hsdS_B(r_B^- m_B^-)$ λ(DE3 [lacI lacUV5-T7 gene 1 ind1 sam7 nin5])	*New England Biolabs*
DH5α	F^- endA1 glnV44 thi-1 recA1 relA1 gyrA96 deoR nupG Φ80d*lacZ*ΔM15 Δ(*lacZYA-argF*)U169, $hsdR17(r_K^- m_K^+)$, λ–	*New England Biolabs* or *Invitrogen*
ER2796*	λ-fhuA2 Δ (lacZ)r1 glnV44 mcr-62 trp-31 dcm-6 zed-501::Tn10 hisG1 argG6 rpsL104 dam-16::Kan xyl-7 mtlA2 metR1 (mcrB-hsd-mrr)114::IS10	*New England Biolabs*
NM1261¥	hsdS ($R^+M^+S^-$) λ(DE3) *tetA cat*	Noreen Murray Lab

**E. coli* ER2796 cells were a gift from Dr Richard Morgan (*New England Biolabs*).

¥*E. coli* NM1261 cells were a gift from Professor Noreen Murray (School of Biology, University of Edinburgh, UK) and are a derivative of NK311. The NM strains were converted to DE lysogens as described by McMahon *et al.*, (2009).

Chemically competent cells

E. coli DH5α cells (*Invitrogen*) were made competent using the $CaCl_2$ method (Sambrook & Russell 2001). 5 mL of LB was inoculated with a single colony of *E. coli* DH5α previously grown on an LB plate. The culture was grown at 37 °C for ~3 hours, with vigorous shaking. When the culture had reached an OD_{600} of 0.4, it was put on ice for 10 minutes. The cells were then harvested at 1520 x *g* for 15 minutes and the subsequent supernatant was discarded. The pellet was resuspended in 3 mL of ice-cold 80 mM $MgCl_2$ 20 mM $CaCl_2$ and centrifuged at 1520 x *g* for 10-15 minutes. This step can be repeated to improve competency. The supernatant was discarded and the pellet was resuspended in 200 µL of ice-cold 100 mM $CaCl_2$. These competent cells were transferred to 50 µL aliquots in pre-chilled Eppendorf tubes and stored on ice. The efficiency of transformation increases four- to six fold during the first 12-24 hours of storage at 4 °C (Sambrook & Russell 2001), and as such, the transformation step can be carried out the next day. If the cells were not to be used within this time-frame, they were mixed with glycerol solution (Sambrook & Russell 2001) in a one to one ratio, frozen in a dry ice/ ethanol bath and then stored at -80 °C.

Transformation of Cells

~3 µL of plasmid DNA (10 ng) was used to transform 50 µL of competent cells. The mixture was incubated on ice for 25 minutes. The cells were treated by heat shock at 42 °C for 45 seconds, and then chilled on ice for 2 minutes. 400 µL of LB or SOC medium was added to the mixture, and the cells were incubated for 1 hour at 37 °C, whilst shaking. A pipette was used to transfer 50-100 µL of these cells onto an LB agar plate containing the appropriate antibiotic (100 µg/ mL carbenicillin). A sterile spreader was used to spread the cells evenly over the plate, which was then incubated, upside down at 37 °C overnight.

Gene Expression

E. coli BL21 (DE3) competent cells were transformed with the target gene, which were first used to grow a small culture in 10 mL of LB broth (supplemented with 100 µg/ mL carbenicillin). Isopropyl-β-D-1-thiogalactopyranoside (IPTG) was added to 1 mM, to induce expression. The cells were incubated at 30 °C for 3 hours, in order to identify whether, and the conditions under which the genes would express. All cell cultures were grown in the presence of an antibiotic (carbenicillin 100 µg/ mL) appropriate to the resistance of the vector, for selection purposes.

A further culture of the competent *E. coli* BL21 (DE3) cells, transformed with the target gene, was grown under the same conditions but on a larger scale (250 mL). This was to check whether the protein was soluble, and was performed by lysing the cells, centrifuging them and then analysing the cell-free extract by SDS-PAGE

After the solubility of the target protein was known, a larger scale expression was carried out by one of two ways, fermentation or induction of 4 to 5 L cell culture in baffle flasks.

Large Scale Induction- Fermentation

The fermentation was carried out by Dr. John White in the Fermentation Suite, School of Chemistry, The University of Edinburgh.

A colony of successfully transformed *E. coli* BL21 (DE3) cells, containing the target gene was picked from an agar plate and used to inoculate 250 mL of LB (supplemented with carbenicillin 100 µg/ mL). This was then incubated at 37 ^{o}C for 5 hours, whilst shaking. The cell culture was then used to inoculate 10 L of LB (supplemented with carbenicillin 100 µg/ mL), which was then fermented in a Bioflo4500 (*New Brunswick Scientific*) fermenter at 37 ^{o}C. When the A_{600} of the culture had reached 0.6, expression was induced by addition of IPTG to 1 mM, the temperature was reduced to 25 ^{o}C and left for 4 hours. Cell pellets were harvested by centrifugation at 7277 x *g* for 10 minutes and stored at -20 ^{o}C for future use. The process would normally yield 4 cell pellets of approximately 6 g.

Baffle flask induction

E. coli BL21(DE3) cells containing the target gene were used to inoculate 100 mL of LB (supplemented with carbenicillin 100 µg/mL). These cells were grown over night at 37°C, whilst shaking. The subsequent saturated cell culture was used to inoculate 3 to 5 L LB (supplemented with carbenicillin 100 µg/mL), in a 1 to 50 ratio. The LB was contained in 2 L baffle flasks, 1 L LB per flask. The cell cultures were then incubated at 37°C, whilst shaking, until the optical density at 600 nm had reached ~0.4. Heterologous gene expression was induced by the addition of IPTG (final concentration of 1 mM) and lowering the temperature to 30°C. The cells were then left for a further 3 hours before harvesting the cells by centrifugation (7277 x *g* for 10 min at 4°C). Cell pellets were stored at -20°C until required.

It was established that with this method, the MS fusion was insoluble. Inducing expression of the MS fusion gene at 20 °C and then leaving the cells overnight (~16 hours) solved this issue, and this method was used for subsequent gene constructs that were presumed to be more fragile.

2.4. Protein Modelling, Purification and Analysis

Genome Searches, Protein Sequence Alignments and Modelling:

All genome searches were performed using *BLAST* (Altschul et al. 1990). All alignments were created using *Clustal Omega* (Sievers et al. 2011) and were edited using *Boxshade* (http://www.ch.embnet.org/software/BOX_form.html). Protein structures were downloaded from the *RCSB Protein Data Bank* (PDB) (Berman et al. 2000) and were edited using *PyMOL* (The PyMOL Molecular Graphics System, 2010).

Protein models were created using Phyre^2 online software (Kelley et al. 2015), and then edited in *PyMOL*. Secondary sequence prediction was performed using *PsiPred* (Jones 1999).

Molecular weights, extinction coefficients and isoelectric points (pI) for proteins were estimated by inputting the predicted amino acid sequence of the target protein into *Scripps* Protein Calculator (http://protcalc.sourceforge.net/)

Protein Purification

All nickel affinity chromatography purifications were conducted using 20 mM sodium phosphate buffer with 500 mM NaCl (pH 7.5) unless otherwise stated. All proteins were expressed with a C-terminal HisTag unless otherwise stated.

E. coli BL21 (DE3) cell pellets were resuspended in 1:10 (g: mL) buffer with 20mM imidazole and a dissolved EDTA-free protease inhibitor tablet (*Roche*). The suspension was then sonicated on ice for ~20 minutes (30 seconds on, 30 seconds off) and centrifuged at 7700 x *g* for 45+ minutes at 4 ^{o}C. The resultant supernatant was then loaded onto a pre-equilibrated Histrap FF crude 5 mL column (*GE Healthcare*), using a peristaltic pump at a flow rate of 100 mL/hr. The flow-through was collected into a Falcon tube (50 mL). The column was then washed with 100 mL 20 mM imidazole buffer and the flow-through was collected. This was followed by an elution of the protein with ~10 mL of 500 mM imidazole buffer, discarding the first 3 mL and collecting the next 6 mL. This was then concentrated to ~4 mL, using a 20 mL 30,000 MWCO Vivaspin (*Sartorius*).

Further purification via size exclusion chromatography was carried out using a Superdex 200 column (*GE Healthcare*) and buffer containing 20 mM Tris-HCl pH 8, 10 mM $MgCl_2$, 500 mM NaCl and 7 mM β-Mercaptoethanol at a flow rate of 10 mL/hr. The subsequent fractions

were analysed by SDS-PAGE using 4-12% gradient Bis-Tris gels (*Invitrogen*), run in NuPAGE® MES SDS Running Buffer (*Invitrogen*), from a 20 X stock solution. Those fractions containing the desired protein were concentrated and stored at -20 ^{o}C in 50% glycerol and 50% size exclusion buffer. The resulting concentration of the samples was obtained via a scan with a UV/Visible spectrometer.

Purifications without a HisTag

As they do not possess a HisTag and therefore could not be purified by nickel affinity, the CC398-1 HsdM and the CCHalfS were purified with an anion exchange step at the start, followed by size exclusion (as described above). The purification of the CCHalfS protein also involved a further step, using a heparin column.

Anion exchange chromatography

The pH of the buffer used in this purification was chosen on the basis of the estimated pI of the protein.

A 0.45 μm sterile filter (*Sartorius*) and 10 mL syringe was used to filter the cell-free extract from ~5 g of lysed cells. The 55 ml DEAE column (*Amersham Pharmacia*) was previously equilibrated with buffer for 3 hours. The sample was loaded at a flow rate of 48 mL/ hour. After the crude extract was completely loaded onto the column, the column was washed with the buffer at a flow rate of 60 mL/ hour for an hour in order to remove any proteins that had not stuck to the column. The flow-through sample, loading wash and the buffer wash were collected and run on a SDS-PAGE gel to check for the absence of the target protein. Protein was eluted using a gradient mixture of 500 mL of 0 to 1.0 M NaCl in 20 mM Tris pH 8.0, 10 mM $MgCl_2$,7 mM β-Mercaptoethanol buffer. The gradient mixture was run overnight at a flow rate of 24 mL/ hour. The elution profile was detected using a UV-Vis spectrometer that recorded the absorbance at 280 nm. 10 μL samples of the fractions corresponding to the various peaks obtained were analysed by SDS-PAGE. Fractions containing the target protein were pooled and the concentration was determined by obtaining the absorbance at 280 nm on a Cary UV spectrometer.

Heparin affinity column Purification

The CCHalfS (without HisTag) truncated protein was purified with a heparin column step. After size exclusion, the protein sample was dialysed against 3 L of buffer (20 mM Tris pH. 7.5, 10 mM $MgCl_2$,7 mM β-Mercaptoethanol). The 120 X 20 mm heparin agarose column (*Sigma Aldrich*) was equilibrated with buffer at 50 mL/ hour for 3 hours. The sample was loaded at a flow rate of 50 mL/ hour. After the sample was loaded onto the column, the column was washed with the buffer at the same flow rate, until all unbound material had passed through the column and the UV absorbance at 280 nm had returned to its baseline. The flow-through sample, loading wash and the buffer wash were collected and run on a SDS-PAGE gel to check for the absence of the target protein. Protein was eluted using a gradient mixture of 500 mL of 0 to 1.0 M NaCl in 20 mM Tris pH 7.5, 10 mM $MgCl_2$,7 mM β-Mercaptoethanol buffer. The gradient mixture was run overnight at a flow rate of 25 mL/ hour. 10 µL samples of the fractions corresponding to the various peaks obtained were analysed by SDS-PAGE. Fractions containing the target protein were pooled and the concentration was determined by obtaining the absorbance at 280 nm on a Cary UV spectrometer.

"Quick" Protein Purification

Due to the apparent degradation of the HsdS subunit over time, a quicker purification and immediate subsequent assay were used. This was started with a nickel affinity chromatography purification as described above. The protein sample was eluted in buffer containing 500 mM imidazole, which was removed by buffer exchange. This was performed using a PD-10 desalting column (*GE Healthcare*). The protein sample was eluted in ~6 mL and was concentrated to 2.5 mL, to be loaded on to the PD-10 column.

The protocol included with the PD-10 column was carried out without any alteration, and was as follows:

1. Cap was removed from the top of the column and the bottom tip was removed with a scalpel. The storage buffer was poured off and discarded.
2. Column was held over empty beaker, to catch flow-through.
3. Column was equilibrated with replacement buffer (20 mM Tris-HCl pH 8, 10 mM $MgCl_2$, 500 mM NaCl and 7 mM β-Mercaptoethanol). Column was filled to the top with buffer, which was then left to drip through the column. This was repeated 4 times and each flow-through was discarded.

4. 2.5 mL of sample was applied to the column with a Pasteur pipette. Samples smaller than 2.5 mL were applied to the column, left to pass through the column and then supplemented with the equilibration buffer, to a final volume of 2.5 mL. Flow-through was discarded.
5. Sample was eluted from the column with 3.5 mL of replacement/ equilibration buffer, and collected into a 6 mL 30,000 MWCO Vivaspin (*Sartorius*).

The Sodium phosphate purification buffer was replaced with the buffer used for size exclusion, as described above. Sample concentration was determined by an A_{280} reading from a Cary50, and the sample was immediately used in an assay to determine presence of DNA cleavage activity. The remaining sample was concentrated in the Vivaspin (*Sartorius*) to an appropriate volume.

Purification of the HsdR Subunit

The HsdR, used in all plasmid and genomic DNA cleavage assays, was purified as described in Roberts *et al* (2013). The *hsdR* gene was ligated into the pRSFDuet vector, and induced with 1 mM IPTG. The protein was produced without a HisTag and purified by anion exchange chromatography, size exclusion chromatography and heparin agarose affinity chromatography.

The *hsdR* gene from N315 (CC5) was amplified and cloned by Dr. John White (School of Chemistry, University of Edinburgh). Purification of the HsdR subunit protein was performed by Mr Laurie Cooper (School of Chemistry, University of Edinburgh).

SDS-PAGE

Protein samples were analysed by SDS-PAGE. 2X SDS loading buffer was added to each sample, which was then heated in a water bath for ~10 minutes at 90 °C. Samples were then centrifuged at 18320 x *g* for 4 minutes and ~5 μL from the top of the sample was loaded on to a NuPAGE® 4-12% Bis-Tris gel (*Invitrogen*). Following electrophoresis, gels were fixed and stained in a solution of methanol:acetic acid:water (3:1:6 by volume) containing 0.1% (w/v) Coomassie Brilliant Blue R250 for at least 30 min. The gels were destained in 30% (v/v) methanol, 10% (v/v) acetic acid.

Gel Densitometry

To give an indication of relative subunit yield, gel densitometry was performed. The free, online software, Image J, was downloaded, and used to visualise images of SDS-PAGE gels (Schneider et al. 2012). This software calculated the relative intensity of selected bands, the results from which were normalised by dividing the given value by the corresponding molecular weight of the protein. The values calculated by this process were compared, to give a ratio of subunit yield (see Appendix D for calculations).

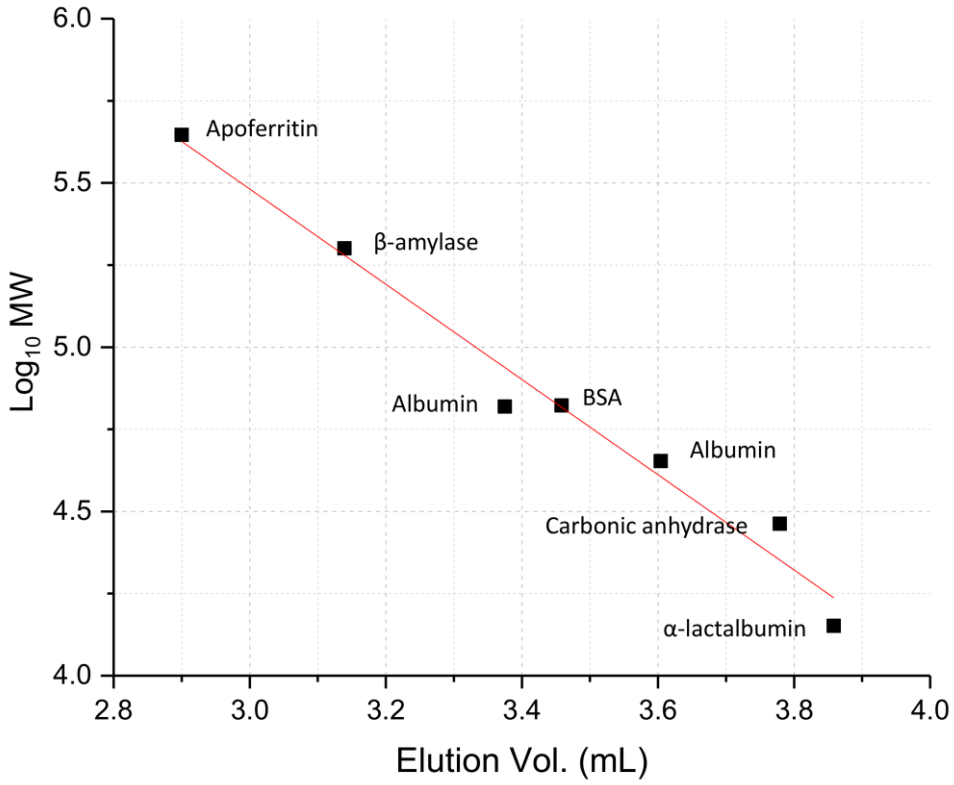
Protein Mass Spectrometry

Protein samples of a range of dilutions were analysed by SDS-PAGE. Measures to achieve highest possible cleanliness were taken, such as clean gloves, fresh running buffer, and the gel tank was cleaned beforehand. Gel bands of appropriate size and density were excised from the gel using a scalpel, and put into a clean, dry Eppendorf tube. These were labelled and sealed with Parafilm® (*Sigma Aldrich*), and then sent for peptide fragmentation mass spectrometry analysis. Analysis was performed using a TripleTOF 5600 electrospray tandem mass spectrometer (*ABSciex*), by Dr Sally Shirran at the BSRC Mass Spectrometry and Proteomics Facility in St. Andrews.

Gel Filtration HPLC

Gel filtration HPLC was carried out as described in Atanasiu *et al* (Nucleic Acids Research, 2002). All runs were carried out using a BioSep-SEC-S 3000 (*Phenomenex*) column and a 20 mM Tris pH 6.5 buffer (20 mM MES, 10 mM $MgCl_2$, 200 mM NaCl, 0.1 mM EDTA, 7 mM β-mercaptoethanol). This buffer was used to dilute the samples to a concentration of approximately 4 µM, 50 µL of which were then injected onto the HPLC system for each run. A flow rate of 0.5 mL/ minute was set for each run, which took approximately 10 minutes to complete. The absorbance at 280 nm was monitored and recorded by a data logger.

First, the column was calibrated using several protein standards (*Sigma Aldrich*) of various concentrations, and a calibration curve was produced (Fig. 19).



	Value	Standard Error
Gradient of the line	-1.44955	0.10229
Y-intercept	9.83011	0.35387

Figure 19: Calibration curve for the BioSep-SEC-S 3000 HPLC column

50 µL of the samples were injected into the 50 µL loop for each run. The rheostat on the HPLC was turned to bring the 50 µL loop into the system and allow the sample to load onto the column. Simultaneously the time logged by the data logger was noted and this was taken as the start of the run. The data gathered from this process was recorded and processed by using Origin (*OriginLab*) software.

MS Fusion Cross-Linking

Several different amounts of MS fusion protein (10, 20, 30, 40 and 50 µg) were used, in order to get the best chance of success. The protein solutions were diluted with Tris buffer, to a volume of 10 mL. To this was added 400 µL 25% (W/V) glutaraldehyde, to give a final concentration of 1% glutaraldehyde. This was incubated at room temperature for 2 minutes, after which the reaction was stopped with 250 µL of fresh 2M $NaBH_4$, dissolved in 0.1M NaOH. This was incubated for 20 minutes in a fume cupboard. After incubation, 10 µL of 10% (W/V) sodium deoxycholate solution (100 mg $C_{24}H_{40}O_4$ in 900 µL ddH_2O).

The protein was then precipitated with 78% (W/V) trichloroacetic acid (TCA) in water. Small amounts of TCA were added incrementally, until the protein precipitate was visible. Approximately 270 µL TCA was added in total. This solution was then split into eight Eppendorf tubes and centrifuged at 18320 x *g* for 10 minutes at room temperature. The supernatant was then removed from these samples, two of which were selected and to one was added 12 µL SDS-PAGE buffer, to the other was added 25 µL. The blue SDS PAGE is turned yellow by the precipitate, but was returned to blue with the addition of ~1 µL of 1M Tris HCl pH 8.8. The samples were then heated at 90 ^{o}C for 10 minutes, left to cool and then centrifuged at 18320 x *g* for 5 minutes. Samples were then analysed by SDS-PAGE.

2.5. Enzyme Assays

Assay to Determine Target Recognition Sites

Assays to determine target recognition sites were carried out as described by Roberts et al., 2013. Briefly, assays were conducted by incubating the enzyme under investigation with a library of plasmids. These plasmids were created by the insertion of known DNA sequences, ligated into the EcoRI- BamHI interval of pUC19. MTases under investigation were supplemented with R subunit from *S. aureus* CC5 and incubated with the plasmid library in separate reactions. Incubations were left for 12 minutes in a water bath set at 37 ^{o}C. The reactions were stopped by the addition of Proteinase K (*Roche*) and incubated in a 60 ^{o}C water bath for 25 minutes. Samples were then analysed by gel electrophoresis.

Enzyme solutions were made up in the following way:

Reagent	Final Concentration (µM)
1 X "Buffer 4" (NEB)	N/A
MTase	3
R' subunit	7.5

Each reaction contained the following:

Reagent	Volume (µL)	Final Concentration
Nuclease-free H_2O	36.75	N/A
10 X Buffer 4 (NEB)	5.00	N/A (1X)
100X BSA (NEB)	0.50	10 µg/µL (1X)
Uncut plasmid	5.00	3.0 nM
SAM	0.75	0.5 nM
ATP	1.00	2 mM
Enzyme	1.00	MTase- 60 nM R'- 150 nM

The place at which Type I R-M systems cleave DNA is not at the recognition sequence. Therefore, a computer program called "RMsearch" (Ellrott et al. 2002) was used to identify the target sequences present in plasmids that had been cut by the enzyme and not present in plasmids that had not.

The CC1-2 MTase was identified as having a recognition site on the pUC19 plasmid, and so every plasmid in the library showed signs of cleavage. This issue was resolved by digestion with 0.5 µL of BamHI (*New England Biolabs*) for 15 minutes at 37 °C, to linearise the plasmid library. The assay was then conducted as described above. Inserts in the plasmid library containing the recognition site for CC1-2 showed a smear on gel electrophoresis.

The plasmids in the library were based on the DNA sequence of phage PhiED1 (a gift from Dr. Garry Blakely, The University of Edinburgh), with inserts from phage λ (a gift from Dr. Iain Murray, *New England Biolabs*). These plasmids are known as "Eddys" and are numbered sequentially from 1E to 20E. N.B. 3E and 8E are not included. The plasmid and specific insert sequences of the Eddys can be found in the supplementary data of Roberts *et al.* 2013.

Assaying RM Fusion Enzymes for DNA Cleavage activity

Assays of restriction activity in the engineered fusion protein were conducted in Eppendorf tubes containing the following:

Reagent	Volume (µL)	Final Concentration
Nuclease-free H_2O	32	N/A
10 X Buffer 4 (NEB)	5	N/A (1X)
100X BSA (NEB)	1	10 µg/µL (1X)
Uncut plasmid	5	3.0 nM
Fusion protein	5	60 nM
SAM (NEB)	2	0.5 nM

± SAM and ± recognition site controls were conducted.

A positive control was conducted in an Eppendorf tube containing the following:

Reagent	Volume (µL)	Final Concentration
Nuclease-free H_2O	35	N/A
10 X Buffer 4 (NEB)	5	N/A (1X)
100X BSA (NEB)	1	10 µg/µL (1X)
Uncut plasmid	5	3.0 nM
CC398-1 protein	1	MTase- 60 nM R'- 150 nM
SAM (NEB)	2	0.5 nM
ATP	1	2 mM

± SAM controls were conducted.

The mixtures were incubated at 37 ºC for 30 minutes, then placed on ice for ≥3 minutes. 2 µL Proteinase K was added and the mixture was then incubated at 60 ºC for 25 minutes. The assay mixture was subsequently analysed by agarose gel electrophoresis.

Agarose gel electrophoresis analyses were conducted using 0.9% agarose gels in 130 mL buffer made from 50 X TAE stock solution. The stock solution was made up as follows: 242 g Tris-HCl, 57.1 mL glacial acetic acid and 100 mL 0.5 M EDTA (pH 8).

Genomic DNA Cleavage Assays

Genomic DNA cleavage assays were conducted by incubating the enzyme with ~0.6 µg of *E. coli* ER2796 genomic DNA or λ phage DNA. MTases were supplemented with R subunit from *S. aureus* CC5 and incubated with the genomic DNA and λ phage DNA, in separate reactions. Incubations were left for 12 minutes in a water bath set at 37 ºC. The reactions were stopped by the addition of Proteinase K and incubated in a 60 ºC water bath for 25 minutes. Samples were then analysed by gel electrophoresis.
Each reaction contained the following:

Reagent	Volume (µL)	Final Concentration
Nuclease-free H_2O	36.75	N/A
10 X Buffer 4	5.00	N/A (1X)
100X BSA (NEB)	0.50	10.0 µg/µL (1X)
Genomic DNA	5.00	1.2 µg/µL
SAM (NEB)	0.75	0.5 nM
ATP	1.00	2 mM
Enzyme	1.00	MTase- 60 nM R- 150 nM

In vivo R-M assays

Virulent unmodified bacteriophage λ (λv.o) were provided by Professor Noreen Murray. The letter "v" indicates that the phage are virulent and unable to form lysogens.

Initially, spot tests were performed to detect the presence of restriction or modification activity in the wild-type, truncated and fusion Sau1 enzymes. The first stage of this procedure involved transforming *E. coli* NM1261 (restriction and modification deficient) cells with a plasmid containing the desired Sau1 gene. The cells were also transformed with the empty pJF plasmid, to act as a control. One colony from each type of the subsequent transformants was then used

to inoculate 5 mL of LB (100 µg/ mL carbenicillin), in a 50 mL Falcon tube. This was left overnight, shaking at 37 ºC. 200 µL of the overnight cultures was mixed with 3 mL of molten BBL "Top" agar (held at 45 ºC) and then poured over the surface of a BBL agar plate (100 µg/mL carbenicillin). This was left for approximately 30 minutes to set, creating ±MTase plates, on which unmodified virulent λ phage was spotted. 5 µL of phage solution was spotted at dilutions from 10^{-1} to 10^{-8}. The plate was left to dry and then placed upside down in an incubator overnight at 37 ºC. The result of this process was plates covered in a bacterial lawn, in which phage plaques had formed. The concentration of the plaques was directly proportional to the concentration of the initial phage concentration. The spot at the dilution at which the plaques were few and separate was ascertained and single plaques were picked from this with a cut pipette tip, and placed in 70 µL phage buffer. 10 µL of chloroform was then added to the buffer and the mix was left at 4 ºC for at least 10 hours. Phage that had been passaged through cells containing the empty plasmid were considered unmodified, whilst phage that had been passaged through cells containing the target gene were potentially modified.

NM1261 cells containing the pJF plasmid with the target gene, were transformed with either an empty pRSFduet plasmid or the plasmid with the CC5 *hsdR* gene. Subsequent transformants were used to produce an overnight culture by the method described previously. 200 µL of this culture was then used to inoculate 3 mL of molten BBL "Top" agar, the mix was poured over a plate of BBL "Bottom" agar (with 100 µg/ mL carbenicillin, 50 µg/ mL kanamycin) and left to set. This produced two types of plate, a + and a – HsdR, on which the retrieved phage was spotted. The phage mix was centrifuged at 18320 x *g* for 3 minutes and used to make dilutions from 10^{-1} to 10^{-5}. 5 µL of each dilution of the unmodified phage, starting from neat solution, was spotted onto the surface of the plate. The plate was left to dry and then placed upside down in an incubator overnight at 37 ºC. The result of this process was a plate, covered in a bacterial lawn, in which phage plaques had formed. If the +R strain of NM1261 cells were exhibiting restriction activity, there would be "cut-back" of the phage plaques. This is to say that there would be at least one order fewer phage plaques on the +R plates, relative to the –R plates. The same process was used to identify modification activity. The potentially modified phage was spotted in the same fashion. Modified phage would show little to no degree of cut-back by the modifying MTase +R.

Efficiency of plating (E.O.P.) assay

To establish the efficiency of plating (E.O.P.) of the modified and unmodified λ phage, whole plate assays were conducted. This aimed to produce plaques across a whole plate, which could be counted by eye. Restriction activity was tested by plaque formation with the +R strain, relative to the –R strain, using unmodified phage. Methylation activity was tested by plaque formation with the +R strain, relative to the –R strain, using modified phage.

The titre at which the modified and unmodified phage produced separate plaques in the spot tests indicated the appropriate dilution of the new stock phage to use in whole plate assays, using the ±R strains. The phage stock was diluted accordingly and mixed with 200 µL of an overnight culture of cells containing the target gene. The mixture was added to 3 mL of molten BBL "Top" agar (held at 45 ^{o}C) and then poured over the surface of a BBL agar plate (supplemented with 100 µg/mL carbenicillin). This was left for approximately 30 minutes to set and then placed upside down and left to incubate overnight at 37 ^{o}C. The following day, the plates were removed from the incubator and the individual plaques were counted. The assay was performed three times, with at least 50 phage plaques per plate. The E.O.P. for the modified and unmodified phage λ was determined using the titre in R proficient strains (NM1261 r+) relative to the titre in R deficient strains (NM1261 r-), and was calculated using the following equation:

E.O.P. = (phage titre of r+m+ strain)/ (phage titre of r-m- strain)

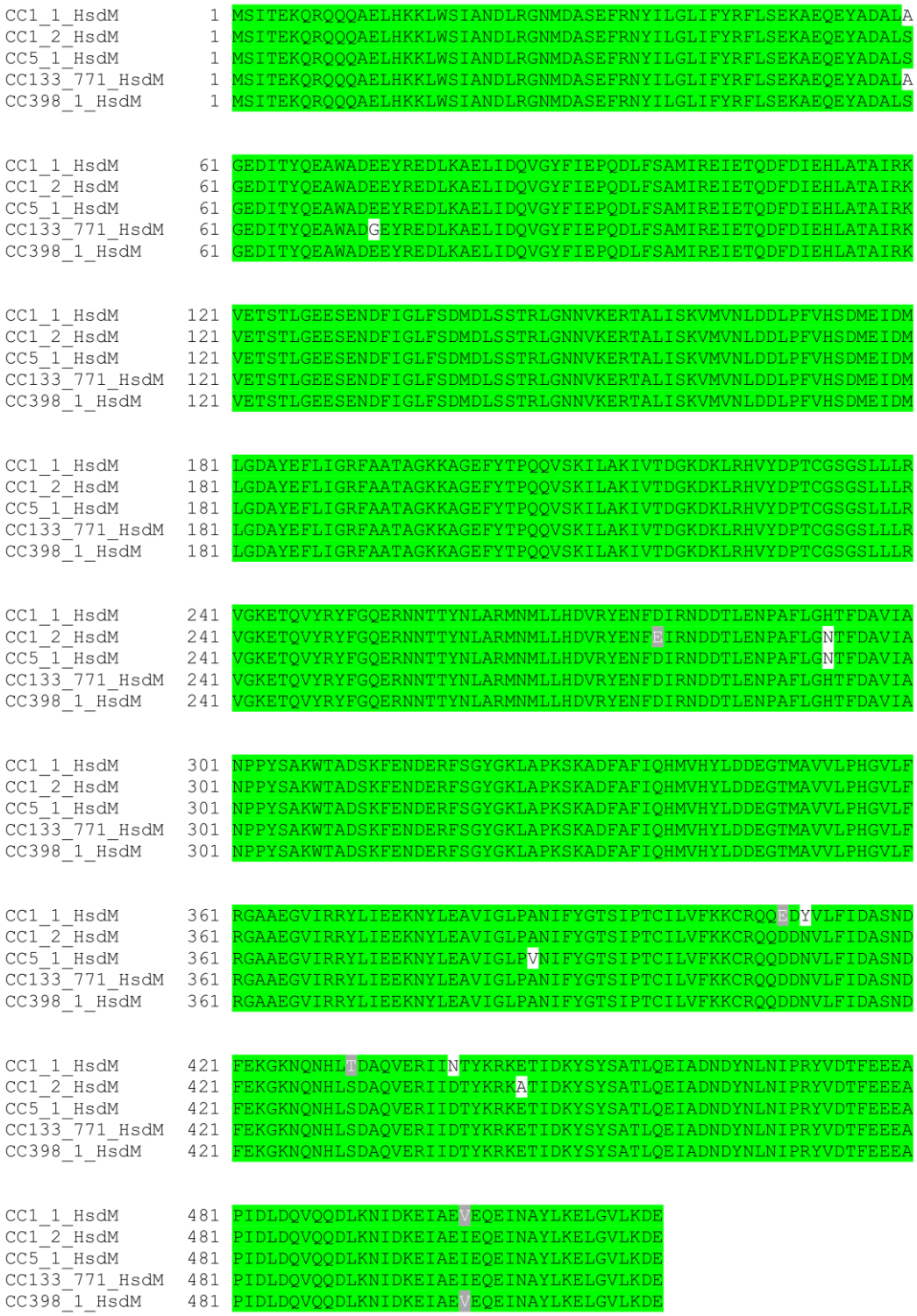
Chapter Three:

Results

3.1. *Staphylococcus aureus* Sau1 Methyltransferases

Several of the Sau1 MTases have been cloned, expressed and the subsequent recombinant proteins purified successfully by the Dryden group (Chen et al. 2010). These recombinant proteins were produced with hexahistidine and GFP tags, from a pJFHisEGFP vector plasmid. With a view to using this plasmid to clone new fusions and other, longer gene sequences, it was decided to remove from it the extra GFP sequence. Although the Sau1 enzymes had been purified and assayed with success, it was also thought that removing this relatively large tag might have a positive effect on the activity of the subsequent recombinant proteins. A thermocycler was used to perform the PCR to eliminate the GFP sequence and produce a vector with a single C-terminal HisTag. This vector was called pJF.

There is a very high degree of conservation between the Sau1 R-M systems from different strains of *Staphylococcus aureus*. In particular, the MTase subunits from these systems only differ to any significant extent in their target recognition domains (TRDs). It is this aspect that gives rise to their difference in DNA sequence recognition. Four MTases from three *S. aureus* lineages (clonal clusters), CC1-1, CC1-2, CC5-1 and CC133-771, had been studied by the Dryden lab for their investigation into R-M systems. Alignment of the amino acid sequences of these MTases (along with that from a CC398 strain, work on which is described later in this thesis), shows the degree of sequence conservation between them (Fig. 20). The almost complete coverage between the HsdM subunits (conserved residues are highlighted in green) and the sequence divergence in the TRDs of the HsdS subunits (conserved residues are highlighted in yellow) is notable. In each case, between the two TRDs is the central conserved region (red outline), which corresponds to the random nucleotides (or N number) in the DNA target. This amino acid spacer sequence is highly conserved across all five examples, as it is across all the Sau1 MTases. The published data from the Dryden lab describes the different DNA recognition sequences of these MTases and specifically, which TRD corresponds to which half of the bipartite sequence (Table 1) (Roberts, et al. 2013). It should also be observed that MTases CC1-1 and CC1-2 (two MTases from CC1) share a high degree of similarity in their TRD1, and therefore half of their DNA recognition sequence is the same.


CC1_1_HsdM 1 MSITEKQRQQQAELHKKLWSIANDLRGNMDASEFRNYILGLIFYRFLSEKAEQEYADALA
CC1_2_HsdM 1 MSITEKQRQQQAELHKKLWSIANDLRGNMDASEFRNYILGLIFYRFLSEKAEQEYADALS
CC5_1_HsdM 1 MSITEKQRQQQAELHKKLWSIANDLRGNMDASEFRNYILGLIFYRFLSEKAEQEYADALS
CC133_771_HsdM 1 MSITEKQRQQQAELHKKLWSIANDLRGNMDASEFRNYILGLIFYRFLSEKAEQEYADALA
CC398_1_HsdM 1 MSITEKQRQQQAELHKKLWSIANDLRGNMDASEFRNYILGLIFYRFLSEKAEQEYADALS
CC1_1_HsdM 61 GEDITYQEAWADEEYREDLKAELIDQVGYFIEPQDLFSAMIREIETQDFDIEHLATAIRK
CC1_2_HsdM 61 GEDITYQEAWADEEYREDLKAELIDQVGYFIEPQDLFSAMIREIETQDFDIEHLATAIRK
CC5_1_HsdM 61 GEDITYQEAWADEEYREDLKAELIDQVGYFIEPQDLFSAMIREIETQDFDIEHLATAIRK
CC133_771_HsdM 61 GEDITYQEAWADGEYREDLKAELIDQVGYFIEPQDLFSAMIREIETQDFDIEHLATAIRK
CC398_1_HsdM 61 GEDITYQEAWADEEYREDLKAELIDQVGYFIEPQDLFSAMIREIETQDFDIEHLATAIRK
CC1_1_HsdM 121 VETSTLGEESENDFIGLFSDMDLSSTRLGNNVKERTALISKVMVNLDDLPFVHSDMEIDM
CC1_2_HsdM 121 VETSTLGEESENDFIGLFSDMDLSSTRLGNNVKERTALISKVMVNLDDLPFVHSDMEIDM
CC5_1_HsdM 121 VETSTLGEESENDFIGLFSDMDLSSTRLGNNVKERTALISKVMVNLDDLPFVHSDMEIDM
CC133_771_HsdM 121 VETSTLGEESENDFIGLFSDMDLSSTRLGNNVKERTALISKVMVNLDDLPFVHSDMEIDM
CC398_1_HsdM 121 VETSTLGEESENDFIGLFSDMDLSSTRLGNNVKERTALISKVMVNLDDLPFVHSDMEIDM
CC1_1_HsdM 181 LGDAYEFLIGRFAATAGKKAGEFYTPQQVSKILAKIVTDGKDKLRHVYDPTCGSGSLLLR
CC1_2_HsdM 181 LGDAYEFLIGRFAATAGKKAGEFYTPQQVSKILAKIVTDGKDKLRHVYDPTCGSGSLLLR
CC5_1_HsdM 181 LGDAYEFLIGRFAATAGKKAGEFYTPQQVSKILAKIVTDGKDKLRHVYDPTCGSGSLLLR
CC133_771_HsdM 181 LGDAYEFLIGRFAATAGKKAGEFYTPQQVSKILAKIVTDGKDKLRHVYDPTCGSGSLLLR
CC398_1_HsdM 181 LGDAYEFLIGRFAATAGKKAGEFYTPQQVSKILAKIVTDGKDKLRHVYDPTCGSGSLLLR
CC1_1_HsdM 241 VGKETQVYRYFGQERNNTTYNLARMNMLLHDVRYENFDIRNDDTLENPAFLGHTFDAVIA
CC1_2_HsdM 241 VGKETQVYRYFGQERNNTTYNLARMNMLLHDVRYENFEIRNDDTLENPAFLGNTFDAVIA
CC5_1_HsdM 241 VGKETQVYRYFGQERNNTTYNLARMNMLLHDVRYENFDIRNDDTLENPAFLGNTFDAVIA
CC133_771_HsdM 241 VGKETQVYRYFGQERNNTTYNLARMNMLLHDVRYENFDIRNDDTLENPAFLGHTFDAVIA
CC398_1_HsdM 241 VGKETQVYRYFGQERNNTTYNLARMNMLLHDVRYENFDIRNDDTLENPAFLGHTFDAVIA
CC1_1_HsdM 301 NPPYSAKWTADSKFENDERFSGYGKLAPKSKADFAFIQHMVHYLDDEGTMAVVLPHGVLF
CC1_2_HsdM 301 NPPYSAKWTADSKFENDERFSGYGKLAPKSKADFAFIQHMVHYLDDEGTMAVVLPHGVLF
CC5_1_HsdM 301 NPPYSAKWTADSKFENDERFSGYGKLAPKSKADFAFIQHMVHYLDDEGTMAVVLPHGVLF
CC133_771_HsdM 301 NPPYSAKWTADSKFENDERFSGYGKLAPKSKADFAFIQHMVHYLDDEGTMAVVLPHGVLF
CC398_1_HsdM 301 NPPYSAKWTADSKFENDERFSGYGKLAPKSKADFAFIQHMVHYLDDEGTMAVVLPHGVLF
CC1_1_HsdM 361 RGAAEGVIRRYLIEEKNYLEAVIGLPANIFYGTSIPTCILVFKKCRQQEDYVLFIDASND
CC1_2_HsdM 361 RGAAEGVIRRYLIEEKNYLEAVIGLPANIFYGTSIPTCILVFKKCRQQDDNVLFIDASND
CC5_1_HsdM 361 RGAAEGVIRRYLIEEKNYLEAVIGLPVNIFYGTSIPTCILVFKKCRQQDDNVLFIDASND
CC133_771_HsdM 361 RGAAEGVIRRYLIEEKNYLEAVIGLPANIFYGTSIPTCILVFKKCRQQDDNVLFIDASND
CC398_1_HsdM 361 RGAAEGVIRRYLIEEKNYLEAVIGLPANIFYGTSIPTCILVFKKCRQQDDNVLFIDASND
CC1_1_HsdM 421 FEKGKNQNHLTDAQVERIINTYKRKETIDKYSYSATLQEIADNDYNLNIPRYVDTFEEEA
CC1_2_HsdM 421 FEKGKNQNHLSDAQVERIIDTYKRKATIDKYSYSATLQEIADNDYNLNIPRYVDTFEEEA
CC5_1_HsdM 421 FEKGKNQNHLSDAQVERIIDTYKRKETIDKYSYSATLQEIADNDYNLNIPRYVDTFEEEA
CC133_771_HsdM 421 FEKGKNQNHLSDAQVERIIDTYKRKETIDKYSYSATLQEIADNDYNLNIPRYVDTFEEEA
CC398_1_HsdM 421 FEKGKNQNHLSDAQVERIIDTYKRKETIDKYSYSATLQEIADNDYNLNIPRYVDTFEEEA
CC1_1_HsdM 481 PIDLDQVQQDLKNIDKEIAEVEQEINAYLKELGVLKDE
CC1_2_HsdM 481 PIDLDQVQQDLKNIDKEIAEIEQEINAYLKELGVLKDE
CC5_1_HsdM 481 PIDLDQVQQDLKNIDKEIAEIEQEINAYLKELGVLKDE
CC133_771_HsdM 481 PIDLDQVQQDLKNIDKEIAEIEQEINAYLKELGVLKDE
CC398_1_HsdM 481 PIDLDQVQQDLKNIDKEIAEVEQEINAYLKELGVLKDE

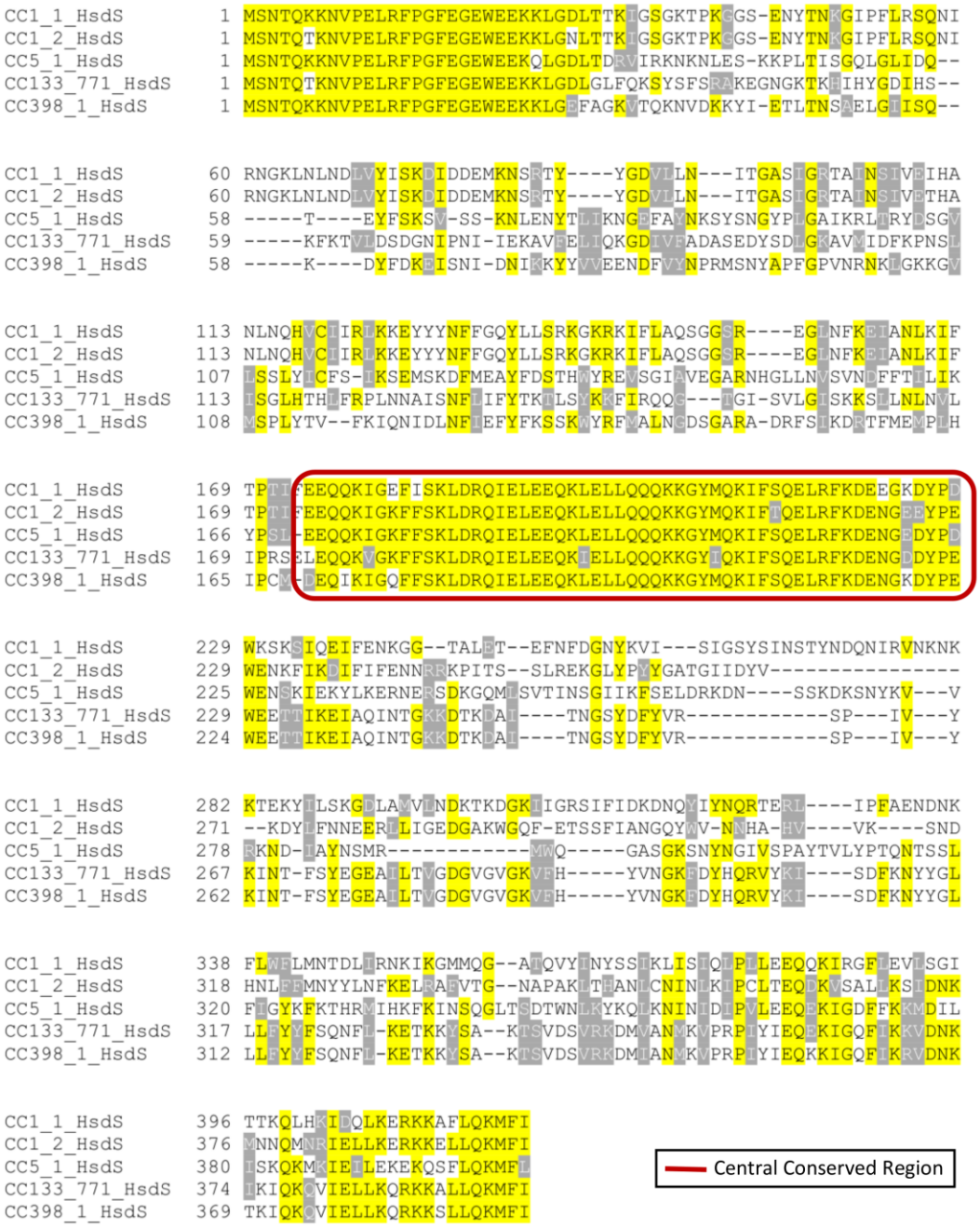


Figure 20: An annotated alignment of the amino acid sequences of five Sau1 HsdM and HsdS subunits from four different strains of *S. aureus*. Conserved residues in the HsdM are highlighted in green, whilst those in the HsdS are highlighted in yellow. Residues with chemically similar side-chains are highlighted in grey. The central conserved region is indicated by the red outline, and clearly shows the high level of conservation between the five enzymes.

HsdS Origin Strain	TRD1	N number	TRD2
CC1-1	CCAY	5	TTAA
CC1-2	CCAY	6	TGT
CC5-1	AGG	5	ATC
CC133/771	CAG	5	RTGA
CC398-1	ACC	5	RTGA

Table 1: The DNA recognition sequences of the separate Sau1 TRDs of five different Type I MTases.

The four sets of Sau1 genes encoding the CC1-1, CC1-2, CC5-1 and CC133-771 MTases, were ligated into vector pJF, which was then used to transform competent *E. coli* DH5-α cells. Plasmid DNA was then purified from cultures of these cells and used to transform competent *E. coli* BL-21 (DE3) cells. The eight (four M, four S) genes were over-expressed and the recombinant proteins were purified successfully (Fig. 21).

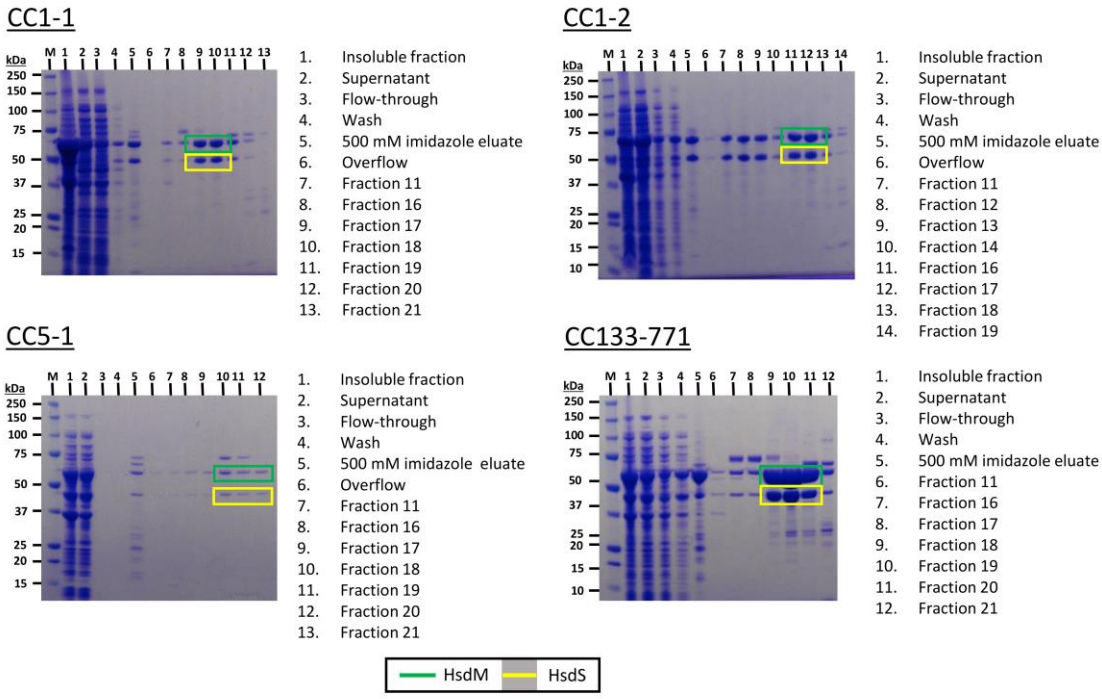


Figure 21: SDS-PAGE gels showing samples from the purification process of four Sau1 MTases.

By observing each protein in solution, it could be seen that the purified Sau1 MTases no longer possessed a GFP tag. Comparisons with the previously purified GFP-tagged proteins showed the complete absence of the characteristic green colour (Fig. 22).

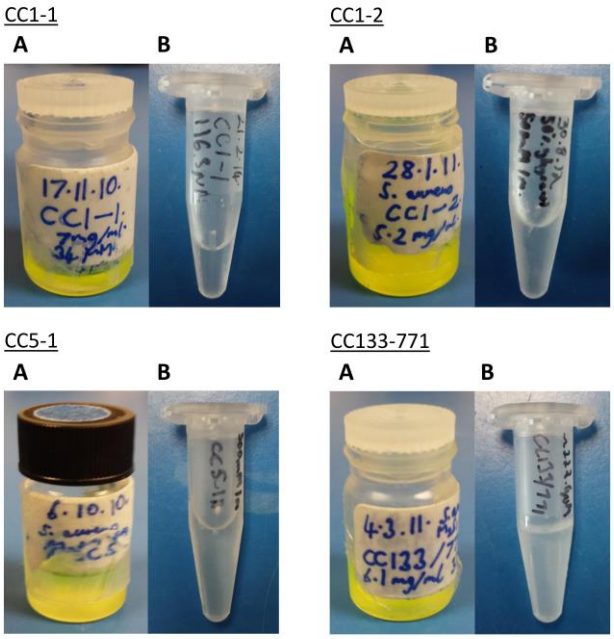


Figure 22: Visual comparison of GFP-tagged (A) to non-tagged (B) recombinant Sau1 MTases.

The proteins were then used in an assay to identify restriction activity. To each MTase was added HsdR subunit protein from the N315 (CC5) strain of *S. aureus*, and these were incubated with a plasmid ("Eddy") library (see Materials and Methods for more details). The results were then visualised on a transilluminator. The pattern of absence and presence of DNA cleavage identifies the recognition sequence of the enzyme, via the computer programme, RMSearch (Ellrott et al. 2002). Each protein was shown to possesses restriction activity and the correct cleavage pattern, as previously determined by the Dryden lab (Figs. 23 to 26) (Roberts, et al. 2013).

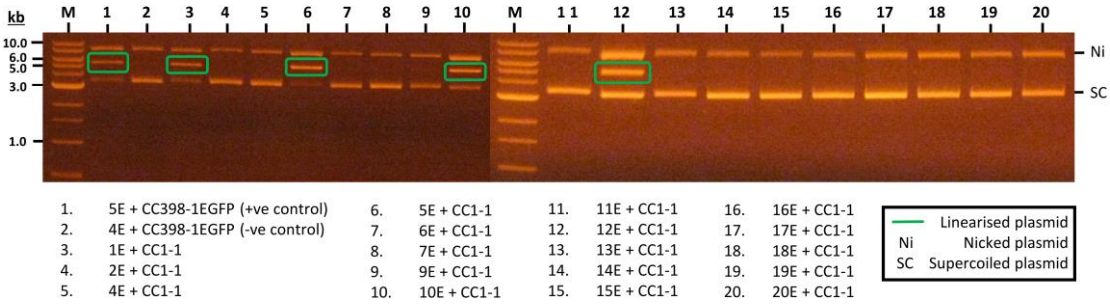


Figure 23: Gel electrophoresis analysis of a plasmid cleavage assay of the CC1-1 MTase. Linearised (cleaved) plasmid species appears on the agarose gel, between supercoiled species (below) and nicked plasmid species (above).

The Sau1 CC1-1 MTase from *S. aureus* CC1 was incubated with the Eddy library. The results showed that it had cleaved Eddys 5E, 10E and 12E (Fig. 23), which is supported by the previously published data from the Dryden lab.

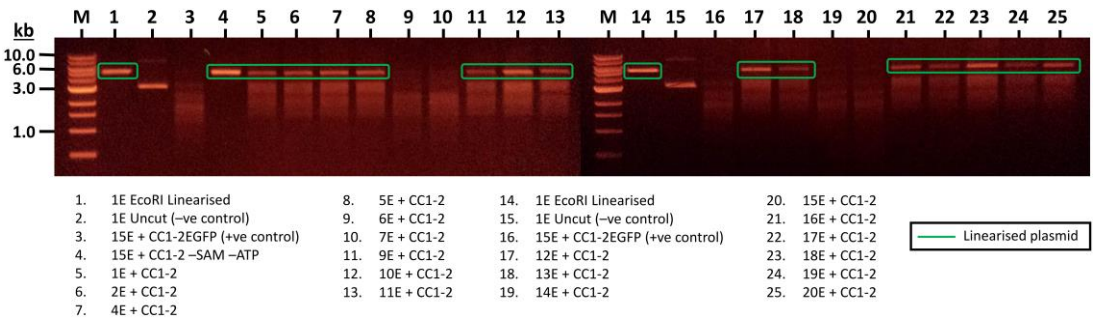


Figure 24: Gel electrophoresis analysis of a plasmid cleavage assay of the CC1-2 MTase.

Previous data from the Dryden lab had shown that the DNA recognition sequence for the Sau1 CC1-2 MTase occurs in the pUC19 plasmid, and so the restriction active CC1-2 MTase would cleave every member of the Eddy library. It had also been discovered that this issue can be resolved by linearising the plasmids beforehand. The CC1-2 MTase was then incubated with the linearised Eddy library. Gel electrophoresis showed that the Eddys had been linearised but that some had been cleaved further and produced a smear. It could be concluded from this that the smears were created by the action of the CC1-2 MTase, and that it had therefore cleaved Eddys 6E, 7E, 14E and 15E (Fig. 24). This result is supported by the previously published data from the Dryden lab.

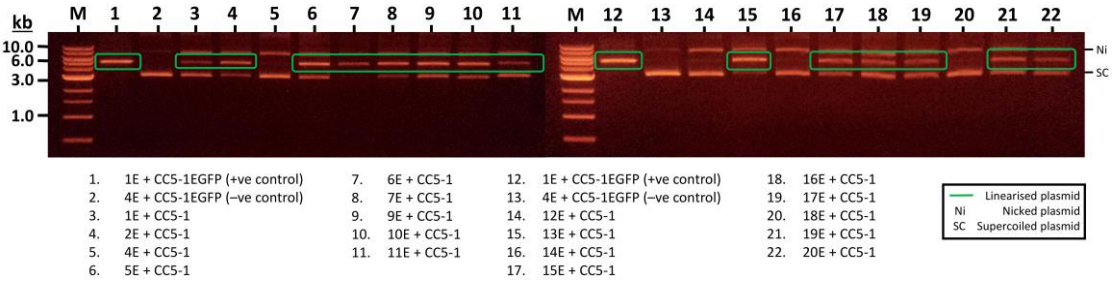
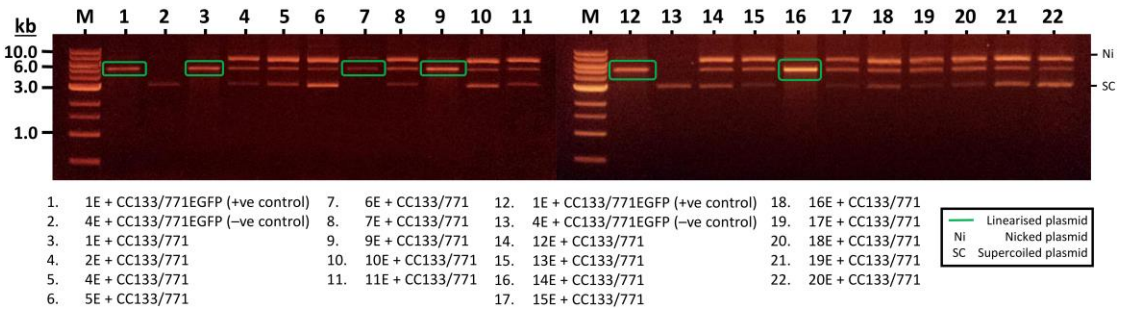


Figure 25: Gel electrophoresis analysis of a plasmid cleavage assay of the CC5-1b MTase.

The Sau1 CC5-1b MTase from *S. aureus* CC5 was incubated with the Eddy library. The results showed that it had cleaved Eddys 1E, 2E, 5E, 6E, 7E, 9E, 10E, 11E, 13E, 15E, 16E 17E, 19E and 20E (Fig. 25). This is corroborated by the previous data from the Dryden lab.



1. 1E + CC133/771EGFP (+ve control)
2. 4E + CC133/771EGFP (–ve control)
3. 1E + CC133/771
4. 2E + CC133/771
5. 4E + CC133/771
6. 5E + CC133/771
7. 6E + CC133/771
8. 7E + CC133/771
9. 9E + CC133/771
10. 10E + CC133/771
11. 11E + CC133/771
12. 1E + CC133/771EGFP (+ve control)
13. 4E + CC133/771EGFP (–ve control)
14. 12E + CC133/771
15. 13E + CC133/771
16. 14E + CC133/771
17. 15E + CC133/771
18. 16E + CC133/771
19. 17E + CC133/771
20. 18E + CC133/771
21. 19E + CC133/771
22. 20E + CC133/771

Linearised plasmid
Ni Nicked plasmid
SC Supercoiled plasmid

Figure 26: Gel electrophoresis analysis of a plasmid cleavage assay of the CC133-771 MTase.

The Sau1 MTase from *S. aureus* CC133-771 was incubated with the Eddy library. The results showed that it had cleaved Eddys 1E, 6E, 9E and 14E (Fig. 26). This too corresponded with the expected results.

Results showed successful purification and assay of the MTases, and that removal of the GFP did not inhibit either aspect. Due to the nature of the assay, it was not possible to get an indication of any effect on reaction rate.

3.2. *Staphylococcus aureus* CC398-1 Methyltransferase

The Sau1 MTase from an *S. aureus* CC398 strain had been purified and assayed successfully by the Dryden lab, but had not been fully characterised. Like the other Sau1 MTases described here, this recombinant protein had originally been purified with an attached GFP tag. This was removed by ligating the gene into vector pJF.

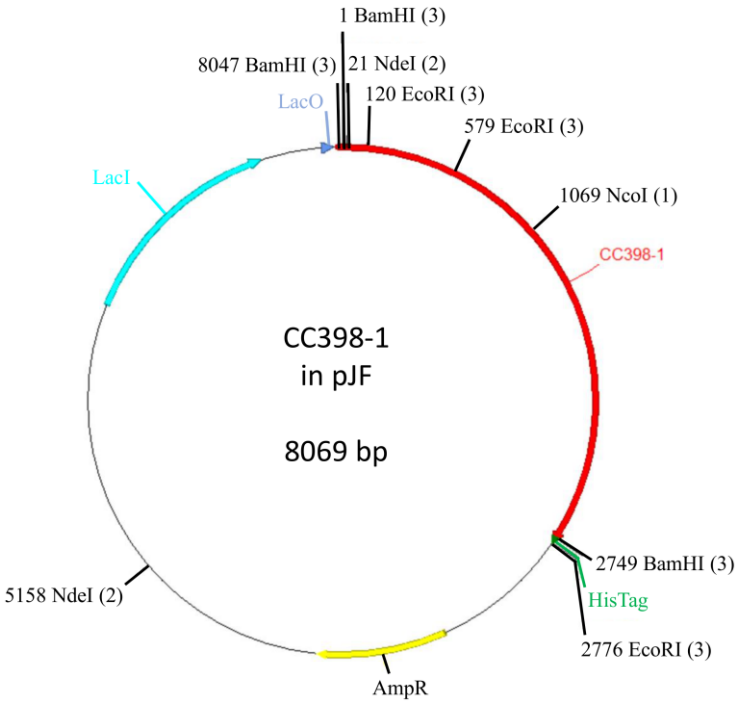


Figure 27: Plasmid map of the CC398-1 MTase genes in vector pJF.

A plasmid map of the CC398-1 MTase in pJF shows how the genes were ligated into the vector, between two BamHI restriction sites (Fig 27). This was the case whenever the pJF vector was used in this work. A potential problem with this cloning strategy is that the genes could be ligated backwards into the vector and would therefore not express. A straightforward screening technique to identify succesful ligation reactions was developed, and involved the culturing of a high number of successful transformants of the ligation reactions, purifying the plasmid DNA from these and then subjecting that to diagnostic BamHI digests. Subsequent analysis by gel electrophoresis identified the presence of an insert in the vector. Selected positives from this stage were sent for gene sequencing, to verify their sequence and determine the orientation of the genes.

A 3D model of the CC398-1 MTase was generated by the free, online software, Phyre^2 (Kelley et al. 2015). This program models the query amino acid sequence against the sequences of

similar proteins of known structure (see Appendix B for matched sequences) . The model of the CC398-1 MTase (M_1S_1 form) shows both constituent HsdM and HsdS subunits (Fig. 28). The two TRDs (highlighted in orange and yellow) of the HsdS subunit can also be observed clearly. The long α-helices separate the two TRDs, and correspond to the spacer sequence (N number) between each part of the bipartite DNA recognition sequence. The results from Phyre[2] stated that the position of 198 amino acids of this model were predicted *ab initio*, and that this is unreliable. However, there was over 90% confidence in 86% of the structure (1234 residues), and so this model does give a very good impression of the structure of the CC398-1 MTase.

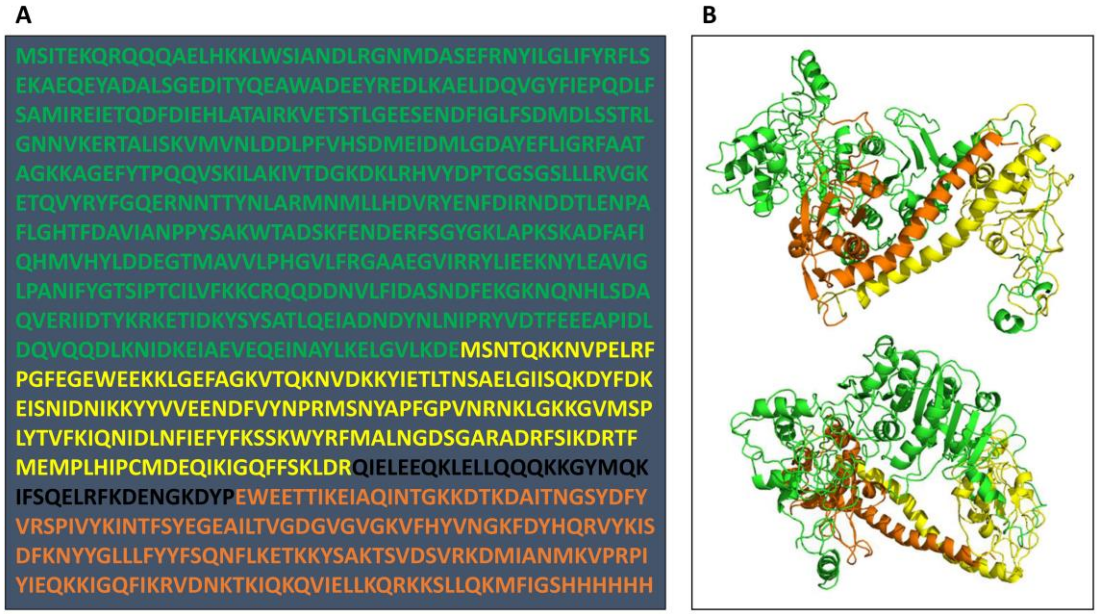


Figure 28: Amino acid sequence of the CC398-1 MTase (A) and 3D model of the protein (B). The HsdM is highlighted in green, TRD1 in yellow and TRD2 in orange.

Positives from the ligation reactions, with the correct gene sequence (as confirmed using FinchTV and ApE software) could then be used for expression tests. One of these was selected and used to transform *E. coli* BL21 (DE3) cells, a single colony of which was then used to create 4 L of culture in LB. Gene expression was induced with IPTG and left for 3 hours. The cells in this culture were then pelleted into two separate aliquots, one of which was resuspended in purification buffer and then sonicated. The sample was then centrifuged and the subsequent cell-free extract was passed over a nickel affinity column, and the target proteins were eluted with imidazole. The eluted protein solution was then purified further by gel filtration chromatography. Samples from selected fractions, along with samples from the nickel affinity steps were analysed by SDS-PAGE to determine the level of purification of the CC398-1 MTase.

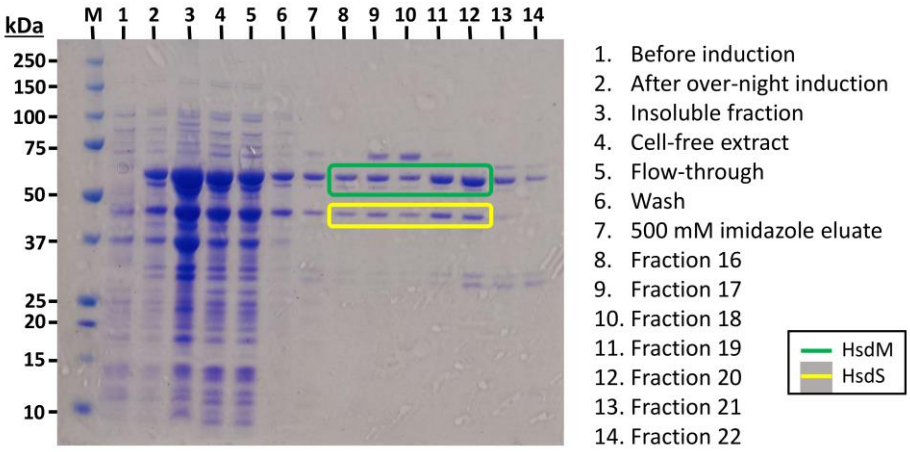


Figure 29: SDS-PAGE analysis of the nickel affinity purification of the CC398-1 MTase.

SDS-PAGE analysis showed the target protein had expressed well and had then been purified (Fig. 29). Protein bands at around the 47 kDa mark (yellow box) corresponded to the HsdS subunit with an additional HisTag. Protein bands at around the 59 kDa mark (green box) corresponded to the HsdM subunit. Gel filtration fractions 17 and 18 appeared to contain a significant level of contaminating protein, whilst fractions 16, 19 and 20 were considered to be of usable purity. These samples of greater purity were pooled and concentrated to approximately 350 µL, and an equal amount of glycerol was added. The final concentration of protein in the solution was calculated to be 20.2 µM, in ~700 µL. The result of this purification was 3.7 mg of the target protein from a 2 L bacterial cell culture. This final solution was analysed by SDS-PAGE to determine its purity (Fig. 30)

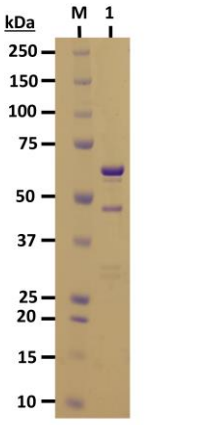


Figure 30: SDS-PAGE analysis
of the purified sample of the CC398-1 MTase.

Active MTase complexes have a M_2S_1 stoichiometry. The HisTag attached to this recombinant protein occurs at the C-terminal of the HsdS subunit and as such, this method of purification relies on the strong affinity between the HsdM and HsdS subunits. It is the S subunit that adheres to the nickel column, and the M is separated from the rest of the cell-free extract by association with the S. Therefore, in order to purify the greatest amount of active MTase, there should be a large amount of HsdS subunit protein and twice as much HsdM protein. The SDS-PAGE gel showed that there was a large amount of HsdM protein (relative to HsdS) but that there seemed to be a relatively low amount of HsdS protein. This would mean a reduced amount of active MTase, but would also be a surprising result given the location of the HisTag. To support this assumption, gel densitometry was performed using the free, online software, Image J (Schneider et al. 2012). An image of an SDS-PAGE gel can be opened in the software, which can then calculate the relative density of selected bands. However, larger proteins bind more Coomassie stain than smaller proteins, and so band intensity is not directly proportional to the amount of the protein. As the bands in question correspond to M and S proteins, which are of different sizes, the values calculated by Image J were normalised by dividing by the molecular weights of the proteins. This process was carried out using four separate sets of the M and S protein bands, the values from which were used to obtain an average. Standard deviation from these values was also calculated, in order to obtain an idea of the margin of error (see Appendix D). The average value for the M subunit was 0.2660 ±0.04, whilst the average value for the S subunit was 0.1270 ±0.05. Normalising these values gave an M:S ratio of 2.09 (±0.31):1 (±0.39). This result indicates that wild-type CC398-1 MTase is being purified successfully, with the correct stoichiometry.

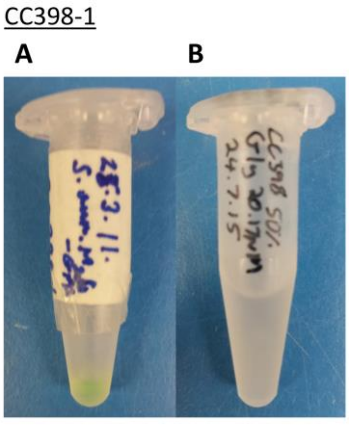


Figure 31: CC98-1 MTase with GFP (A) and without GFP (B).

By comparing this MTase protein solution with that of the recombinant, GFP-tagged CC398-1 MTase, it was observed that the new CC398-1 MTase did not possess the GFP tag (Fig. 31). Gel filtration HPLC was used to provide further information about the stoichiometry of the MTase and the purity of the protein solution (Fig. 32).

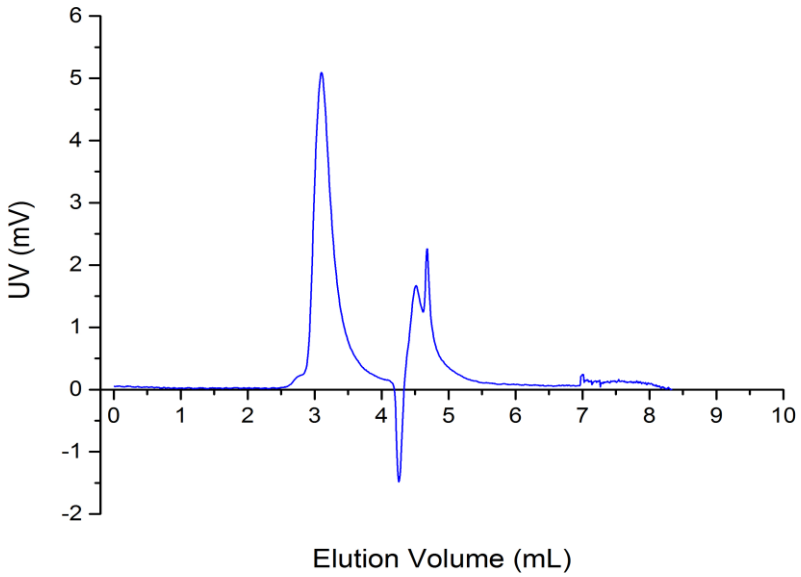


Figure 32: Graph of UV absorbance at 280 nm, against elution volume. Data taken from gel filtration HPLC of 4 µM CC398-1 MTase.

Using a calibration curve (see Materials and Methods) and the elution volume (~3.1 mL) of the MTase from the HPLC column, an estimate of the molecular weight of the protein can be calculated. Carrying out this process gives a result of ~210 kDa. This value does not agree with the expected molecular weight of the CC398-1 MTase is ~166 kDa. Previous work in the Dryden lab has shown that the complete Sau1 MTases do not pass through the HPLC column in the manner that is expected, and so the exact figure calculated for the molecular weight is not reliable. However, the data retrieved from the process can be used qualitatively, and compared with data from other MTases. The HPLC trace from the CC398-1 MTase indicated that the protein solution was relatively pure, as it showed one neat and sharp peak corresponding to the target protein. After this peak, there was a sudden decrease in absorbance, leading to a second peak at an elution volume of around 4.5 mL. This was due to a change in the refractive index, caused by the glycerol in the protein solution. The trace for this wild-type enzyme will be used as the control for the other HPLC work shown in this thesis.

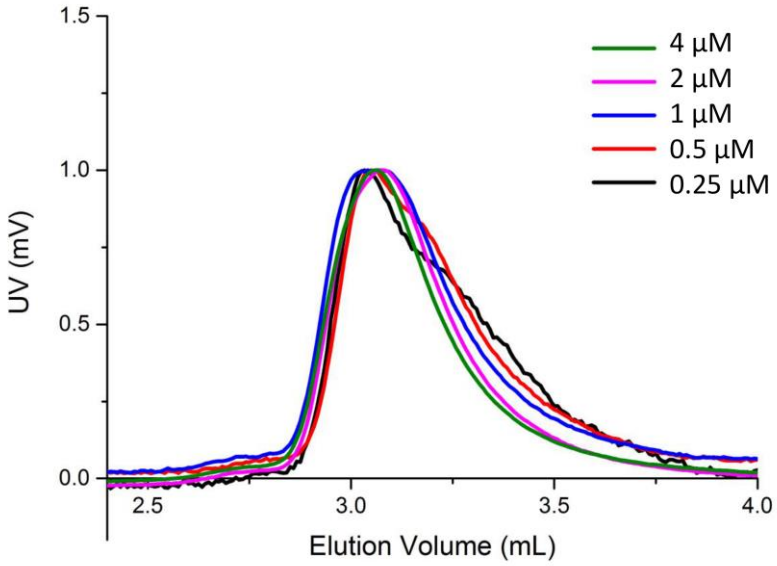


Figure 33: Normalised elution profiles of decreasing concentrations of the CC398-1 MTase.

By comparing the elution profiles of decreasing concentrations of the CC398-1 MTase, the stability of the enzyme's quaternary structure can be observed. The concentration of the protein solutions was halved successively (starting from 4 µM), and analysed by gel filtration HPLC (Fig. 33). A peak corresponding to the enzyme in its natural state occurred at an elution volume of approximately 3.1 mL, as observed previously. With decreasing concentration, a shoulder on the right side of the peak begins to form. This occurs when an M subunit dissociates from the M_2S_1 complex, resulting in the M_1S_1 inactive form. The preparation of protein solution used in this investigation was reliably pure. It could be seen that this wild-type enzyme complex was remarkably stable under the experimental conditions (Tris pH. 6.5), until the curve shifts at a concentration of 0.25 µM and below. As the signal produced by this low concentration is more susceptible to the background signal of the system, it is difficult to determine whether the curve's shift is due solely to the loss of the second HsdM. However, it could be stated that the enzyme complex was stable to at least a concentration of 0.25 µM.

To assess the activity of the CC398-1 MTase, a restriction assay with the Eddy library was carried out. Samples from this assay were analysed by agarose gel electrophoresis and visualised by UV light. The agarose gel showed a clear and precise pattern of cleavage (Eddys 5E, 6E, 7E, 12E and 14E) that corresponded to the data previously obtained by the Dryden lab (Fig. 34). This pattern was used by RMSearch to interpret the DNA target site, ACCNNNNNRTGA.

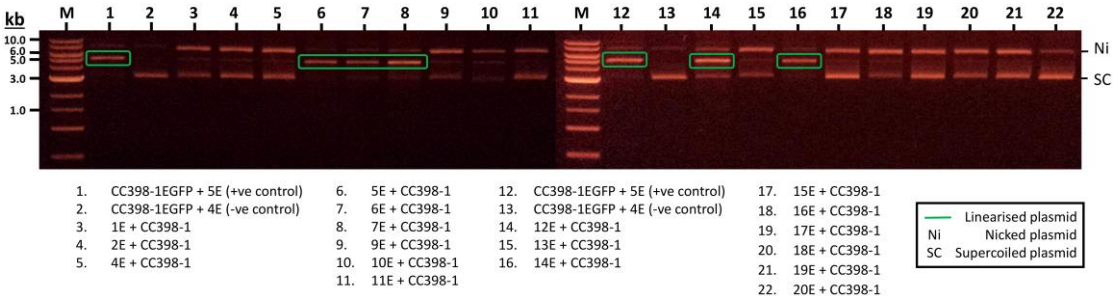


Figure 34: Gel electrophoresis analysis of a plasmid cleavage assay of CC398-1 MTase.

This result showed that the CC398-1 MTase, without EGFP, had restriction activity *in vitro*. *In vivo* studies were used to identify whether this MTase also exhibited modification activity (Fig. 35). The CC398-1 MTase gene in pJF vector was used to transform *E. coli* NM1261 cells. These cells were then transformed with either the empty pRSFDuet plasmid (-R), or the pRSFDuet plasmid containing the CC5 *hsdR* (+R).

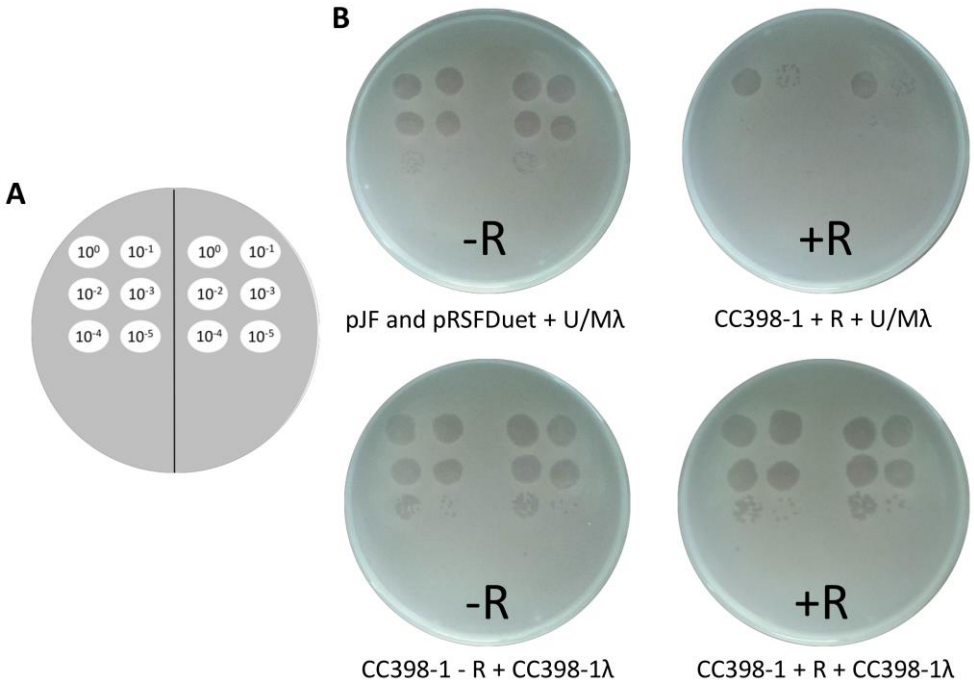


Figure 35: Diagram of spot test dilutions (A). *In vivo* spot test assay of the CC398-1 MTase (B). "-R" plates describe bacteria lacking the *hsdR* subunit gene, whilst those with "+R" do possess it. pJF is an empty plasmid, which would otherwise carry the MTase genes. pRSFDuet is an empty plasmid, which would otherwise carry the *hsdR* gene. U/Mλ denotes phage that are unmodified. CC398-1λ denotes phage that have survived bacteria containing the CC398-1 MTase genes, and are therefore expected to possess the pattern of DNA methylation specific to the CC398-1 MTase.

When these cells also contained the gene for the CC5 HsdR subunit, they showed evidence of restriction activity against unmodified λ phage. Unmodified λ phage was used to infect NM1261 cells that contained the CC398-1 MTase, and the subsequent phage plaques were used infect NM1261 cells that contained both the CC398-1 MTase and the CC5 R genes. In

this instance, there was little evidence of restriction activity. It can be inferred from this result that specific sites along the λ phage DNA had been modified by the CC398-1 MTase, and that these sites were therefore protected from restriction by the restriction complex containing this CC398-1 MTase.

Spot tests indicated that the CC398-1 MTase was modifying λ phage with what appeared to be complete efficiency (E.O.P.= ≥1), and could restrict the phage by at least 3 orders (from 4 to 1 full spots of phage plaques). However, given the difficulty inherent in resolving plaques of such a small size, more accurate data was obtained from whole plate assays. Several rounds of whole plate assays were conducted and the phage plaques were counted (Table 2).

CC398-1 MTase *In vivo* Assay Results:

Phage Type	Phage Dilution	Phage Volume	R-M System	Number of Plaques
Unmodifiedλ	10^{-5}	100 µL	None	1516
	10^{-3}	40 µL	CC398-1 + R	910
	10^{-6}	100 µL	None	263
	10^{-4}	80 µL	CC398-1 + R	644
	10^{-6}	100 µL	None	238
	10^{-4}	50 µL	CC398-1 + R	1898
	10^{-6}	100 µL	None	219
	10^{-5}	50 µL	CC398-1 + R	256
	10^{-6}	100 µL	None	208
	10^{-5}	50 µL	CC398-1 + R	314
398-1λ	10^{-6}	50 µL	CC398-1	80
	10^{-6}	50 µL	CC398-1 + R	79
	10^{-6}	50 µL	CC398-1	74
	10^{-6}	50 µL	CC398-1 + R	70
	10^{-6}	50 µL	CC398-1	135
	10^{-6}	50 µL	CC398-1 + R	89
	10^{-6}	50 µL	CC398-1	90
	10^{-6}	50 µL	CC398-1 + R	96

Table 2: Raw data collected from several rounds of full plate *in vivo* assays of the CC398-1 MTase. The table shows ±R pairs, which are the results from the experiment (+R) and control (-R), and the subsequent repeats. The plaque numbers cannot be compared without adjusting for volume and dilution.

The E.O.P. of the unmodified λ phage against the restriction active 398-1 MTase was 0.15 (see Appendix E for calculations). This value shows that the restriction complex is active. However, the high degree of error (±0.11) highlights that there is significant variability, and so a measure of activity is not possible. The E.O.P. of the 398-1 modified λ phage was 0.88. This agrees nicely with the value estimated from the spot tests of ≥1. It indicates that this is a prolific MTase, as it is modifying the phage such that most are surviving restriction. Nevertheless, there is also a high degree of uncertainty (±0.27) for this value.

3.3. CC398-1 HsdM

An HsdM to HsdS fusion protein had been created by the Dryden lab, using genes from EcoKI (Roberts et al., 2012). This fusion showed both methylation and restriction activity *in vivo*, but to elicit activity *in vitro*, stoichiometric amounts of purified HsdM subunit protein needed to be added. This was in order to establish the M_2S_1 active complex. It was assumed that an MS fusion of the CC398-1 MTase (work on which is presented later in this thesis) would also need to be supplemented with its HsdM subunit. The CC398-1 HsdM would therefore need to be purified individually.

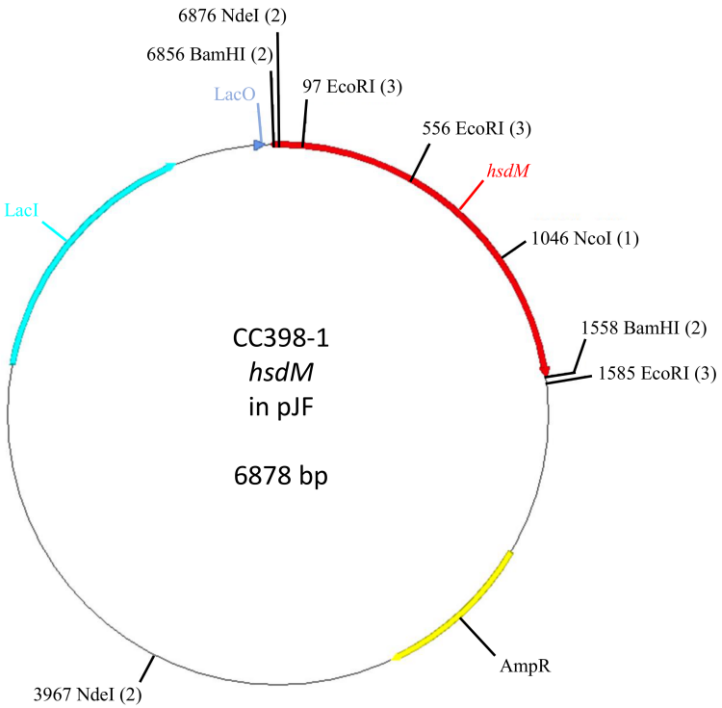


Figure 36: Plasmid map of the CC398-1 *hsdM* in vector pJF.

A plasmid map of the CC398-1 *hsdM* shows the gene ligated into the vector pJF in the manner previously described (Fig. 36). When observing the map, the absence of a HisTag sequence at the 3' of the *hsdM* gene should be noted. The lack of a HisTag on the protein product of this gene means that it is different from the other proteins described in this work. Strictly speaking, the HisTag sequence has been retained on the plasmid, but a stop codon introduced immediately upsteam prevents its transcription. The HsdM subunit was to be used to supplement an MS fusion protein, a process which could be affected by the presence of a HisTag. The *hsdM* gene would express protein without a tag. This could not therefore be purified by nickel affinity chromatography.

The *hsdM* sequence was verified by Sanger sequencing and the plasmid was then used to transform *E. coli* BL21 (DE3) cells. A small scale induction (250 mL LB) was carried out to determine the solubility of the HsdM protein. The subsequent cell culture was sonicated and the cell-free extract was analysed by SDS-PAGE (Fig. 37).

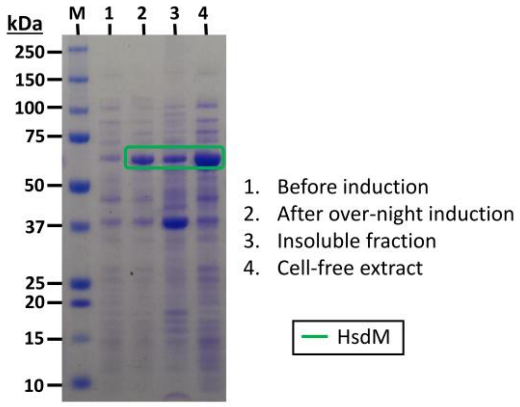


Figure 37: SDS-PAGE analysis of samples from a small scale induction of the *hsdM* gene.

The SDS-PAGE gel showed that the *hsdM* gene had expressed. A strong band at around 59 kDa in the cell-free extract sample corresponded to the size of the HsdM, and indicated that the protein was largely soluble under these conditions. On the basis of this result, a larger scale induction was performed, in order to purify the HsdM protein. The estimated pI of the HsdM protein was 4.79, and so the cell-free extract from 2 L *E. coli* cell culture was passed through an anion exchange column. The samples from the subsequent fractions were subjected to SDS-PAGE analysis (Fig. 38)

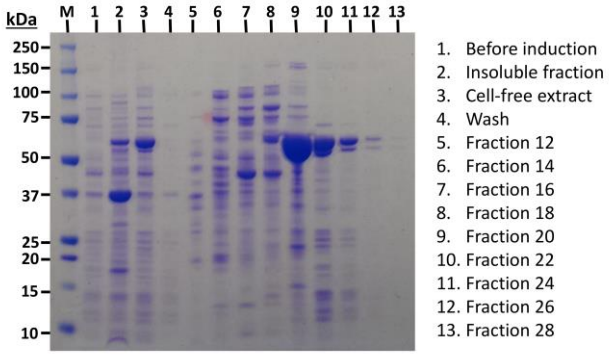


Figure 38: SDS-PAGE analysis of samples from the anion exchange purification of the HsdM protein.

Large bands at the 59 kDa mark on the SDS-PAGE gel indicated that the HsdM protein began to elute from the column at a NaCl concentration of 240 mM. The gel showed that the majority

of HsdM protein was contained in Fractions 20 to 24. Fractions 19 to 22 were pooled and subjected to purification by size exclusion. Samples from the size exclusion step were analysed by SDS-PAGE (Fig. 39).

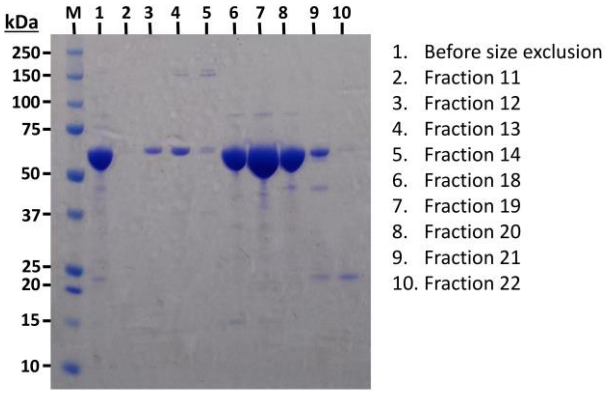


Figure 39: SDS-PAGE analysis of samples
from the size exclusion purification of the HsdM protein.

SDS-PAGE analysis of fractions from the size exclusion purification showed a large amount of relatively pure protein was contained in Fractions 18, 19 and 20. These fractions were pooled and concentrated, and the concentration of the solution was estimated by UV/vis spectroscopy. From this purification of 2 L of cell culture, the estimated yield of HsdM protein was 20.6 mg.

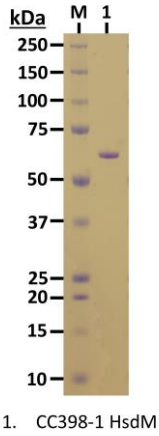


Figure 40: Purified sample of HsdM.

The mixed sample was then subjected to SDS-PAGE to check the relative purity of the HsdM protein (Fig. 40). The sample showed very little sign of contamination, and as such the calculated yield of HsdM was reliable.

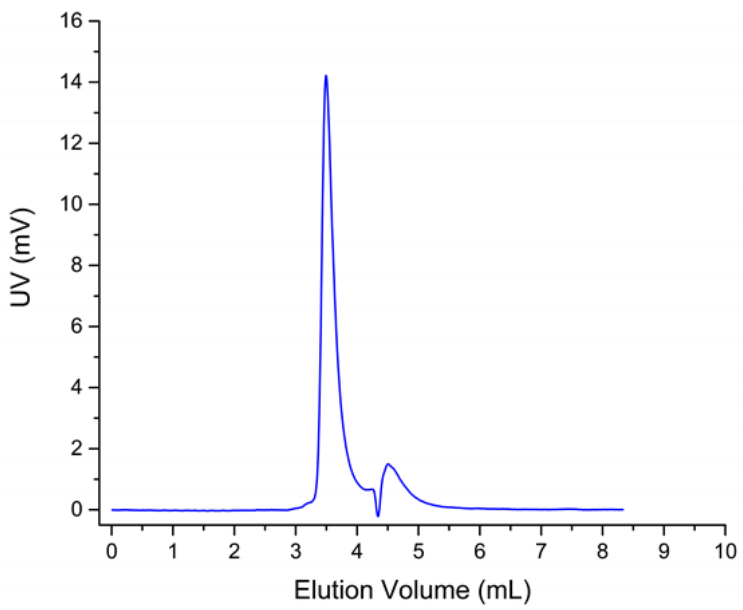


Figure 41: Gel filtration HPLC trace of the HsdM protein.

The gel filtration HPLC data (Fig. 41) supports the other work on the HsdM protein. A single, tight peak on the trace suggests the protein solution was relatively pure. The elution volume of the HsdM protein was 3.48 mL. This value can be used to calculate a molecular weight of ~60.3 kDa, which is close to the expected molecular weight of the protein (59.4 kDa). This also gave confidence that this experiment was providing reliable results.

Having successfully purified the CC398-1 HsdM, a Phyre[2] model was created of the protein, in order to get an impression of its structure (Fig. 42).

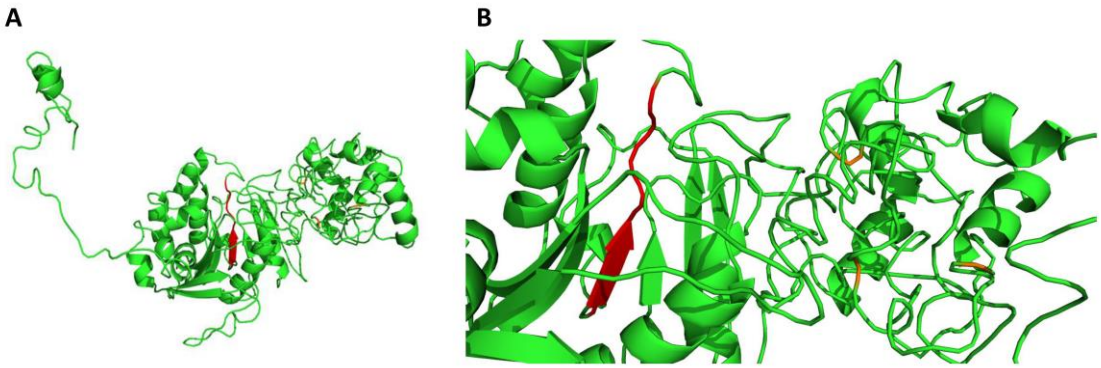


Figure 42: Model of the CC398-1 HsdM (A). A close-up image of the protein (B). The active site (AVIANPPY) is highlighted in red, and the residues of the SAM binding motif (FXGXG) are highlighted in orange.

The model showed the SAM binding motif towards the N-terminal of the protein, co-ordinated by F171, G182 and G190. These residues make up the triangular SAM binding pocket in the protein, shown in orange. Towards the C-terminal is the MTase active site, coloured red. P302, P303 and Y304 are the key catalytic residues, and form a β-strand and coil, amongst a largely α-helical domain. The model also possesses a long C-terminal tail. In the model of the M_1S_1 structure, this tail interacts with the first TRD of the HsdS. This is perhaps how the TRDs are stabilised by the HsdM.

3.4. CC5 HsdR to CC398-1 MTase Fusion

Kennaway *et al.* (2012) proposed that there is an evolutionary link between Type I and Type II R-M systems. That the types share important characteristics has led to the suggestion that removing the motor domains from a Type I system would create a pseudo-Type II system, similar to the structures of the Type IIG and IIB system subtypes. The difficulty with testing this theory is in estimating to where in the amino acid sequence the Type I HsdR subunit should be truncated. The sequence needed to be altered in order to disable its DNA translocation activity but retain its nuclease activity.

As there is not yet a crystal structure of the Sau1 HsdR subunit from the N315 (CC5) strain of *S. aureus*, a model was created using Phyre[2] online software, using the EcoR124I HsdR as a template (Fig. 43) (See Appendix B for further details). This was used in conjunction with results from an online secondary sequence prediction software called *PsiPred* (Jones 1999), to identify appropriate regions to which the protein could be shortened (Fig. 44).

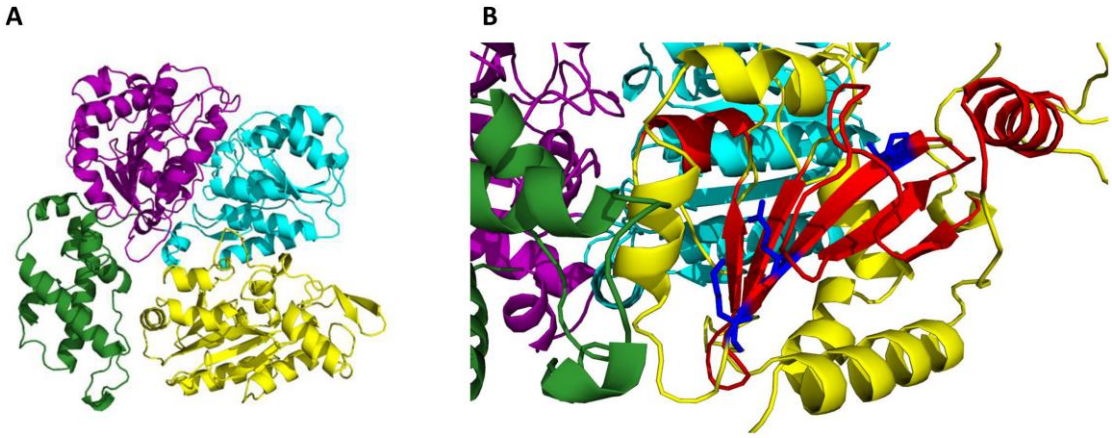


Figure 43: Model of the CC5 Sau1 HsdR Subunit (A), with the nuclease domain highlighted in yellow, the two motor domains in cyan and purple, and the helical domain coloured green. A close-up view of the nuclease active site (B). The conserved α-helical and β-strand structural motif is coloured red, with the key Pro-Asp-Lys catalytic motif in blue.

The model of the CC5 HsdR gave a good impression of the domain organisation of the protein. It indicated that in order to create a successful R to M fusion, the nuclease (yellow) region of the HsdR should be retained. However, it is unclear whether truncating the protein to this point would result in the correct folding of the protein. With this in mind, sites to which the enzyme would be truncated were deemed appropriate on the basis of maintaining its predicted secondary structure. In simple terms, these peptide sequences were shortened, up to the end of an α-helix or a β-strand.

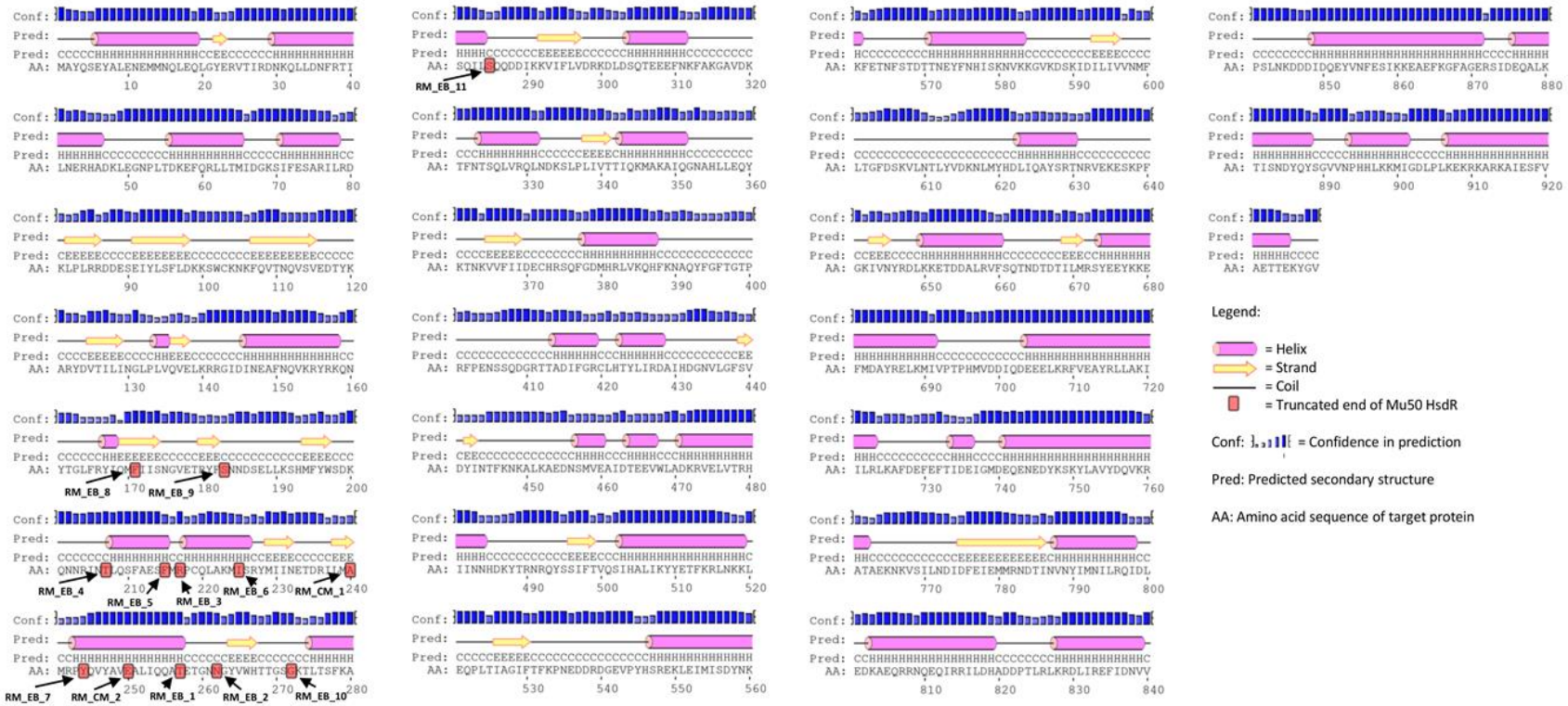


Figure 44: Annotated secondary sequence prediction for the CC5 HsdR subunit.
The sequence prediction was created using the online software, *Psipred* (Jones 1999).

Over the course of this project, a total of 13 Δ*hsdR* to *hsdM* gene fusion constructs were made. A cartoon representation of the different sequence lengths was drawn to give a basic impression of the fusion proteins, and their relative difference in peptide sequence length (Fig. 45). From this diagram, it can be seen that the R subunit (Red) is joined directly to the M subunit (Green), and that the S subunit (Yellow) is not connected to this fusion. This is because, as with all the other wild-type Sau1 RM systems, the HsdS protein is translated as a separate peptide.

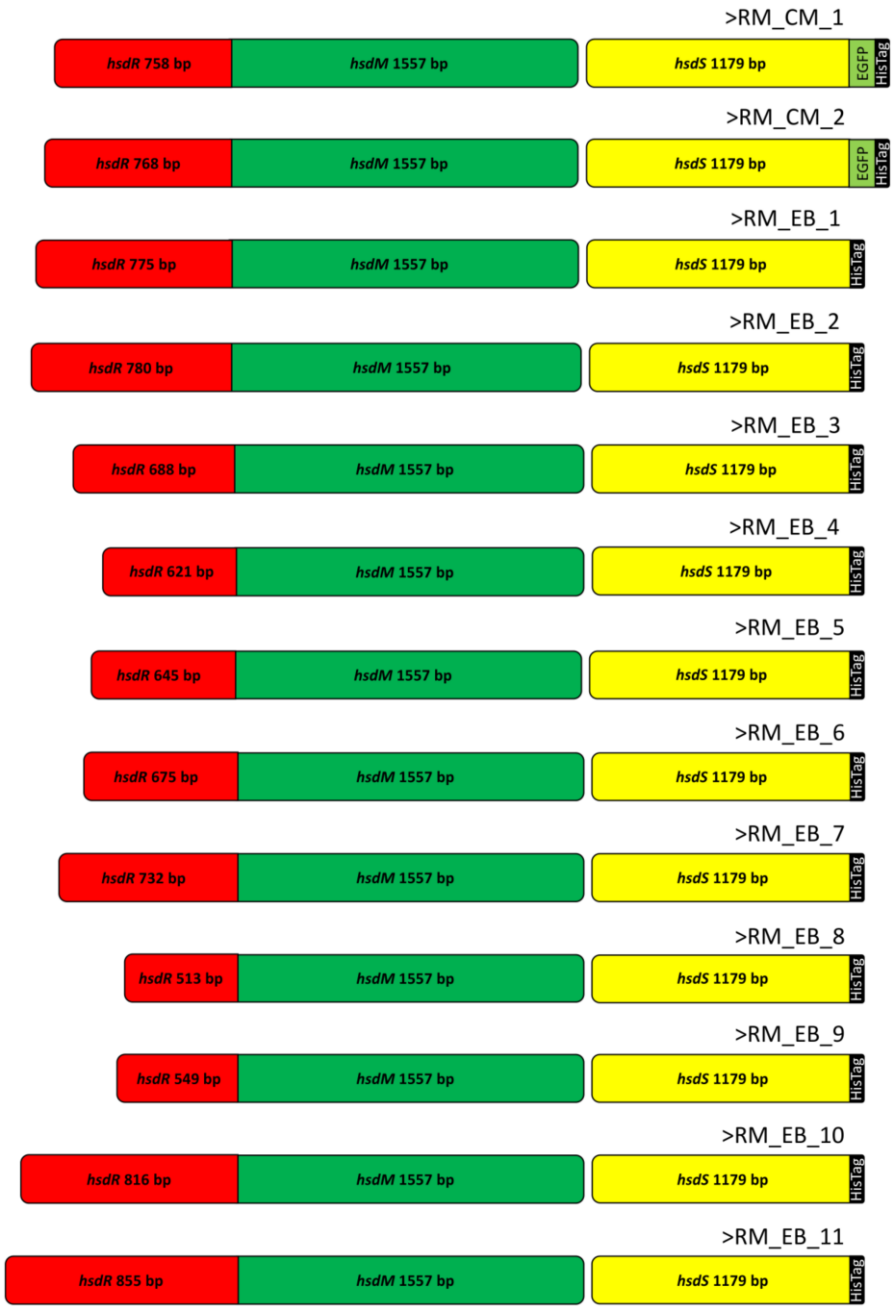


Figure 45: Diagram to show the relative sizes of the R to M fusion genes. The title "RM" denotes an HsdR to HsdM fusion. "CM" denoted that the fusion was created by Chris McLean. "EB" denotes that the fusion was created by Edward Bower.

Two of the RM fusions, RM_CM_1 and RM_CM_2 (2315 and 2325 bp fusion genes respectively), were created at the beginning of this study. All work on these fusions was conducted by Chris McLean but is presented here to give a complete view of the project so far.

The RM_CM_1 and RM_CM_2 genes were ligated into vector pJFMSEGFP. The product of this reaction was used to transform competent *E. coli* DH5-α cells. The plasmid DNA was purified from these cells and then used to transform competent *E. coli* BL-21 (DE3) cells. A small scale IPTG induction of these cells was conducted to test whether the fusion genes would be expressed. 1 mL samples from these cells were taken before induction, 3 hours after induction and 22 hours after induction. These samples were then subjected to analysis by SDS-PAGE (Fig. 46).

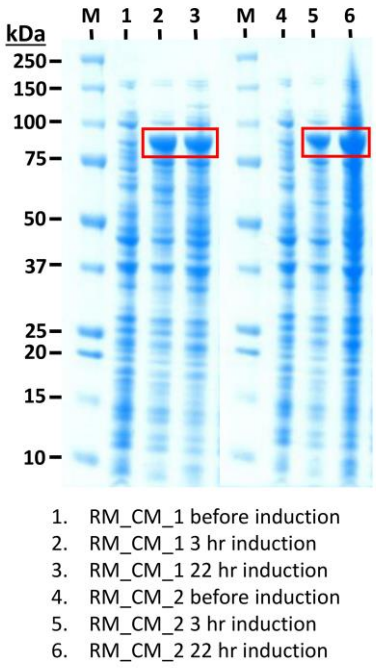


Figure 46: SDS-PAGE analysis of the small scale induction of the RM_CM_1 and 2 genes. The red boxes indicate the presence of over-expressed protein of the expected molecular weight of RM_CM_1 and 2.

The results of this analysis indicated a successful expression of the two different fusion genes. On this evidence, a large scale induction by fermentation was carried out, in order to produce a large amount of the fusion protein. The subsequent recombinant protein possessed GFP and hexahistidine tags, which resulted in protein with a bright green colour that could be purified by nickel affinity chromatography. The cell pellets from the fermentations were resuspended

and lysed by sonication, and the subsequent cell-free extract was purified with a nickel affinity column. The results of this step were run on an SDS-PAGE gel (Fig. 47).

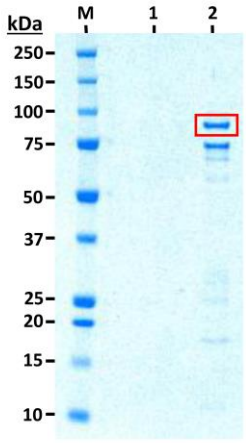


Figure 47: SDS-PAGE analysis of the final eluates from the nickel affinity purification of the first two RM fusion proteins.

SDS-PAGE analysis showed a significant yield of only one of the two proteins (RM_CM_2). To confirm this result, and to check for restriction activity in RM_CM_2, the two protein solutions were used in a DNA cleavage assay. As the fusions contained the complete CC398-1 MTase, it could be assumed that the protein products would have the same DNA recognition sequence as the wild-type MTase. Given this, the protein solutions from the purification of RM_CM_1 and RM_CM_2 were incubated with a known positive (plasmid 7E) and a known negative (plasmid 15E). The results of this assay were run on an agarose gel and visualised on a transilluminator (Fig. 48).

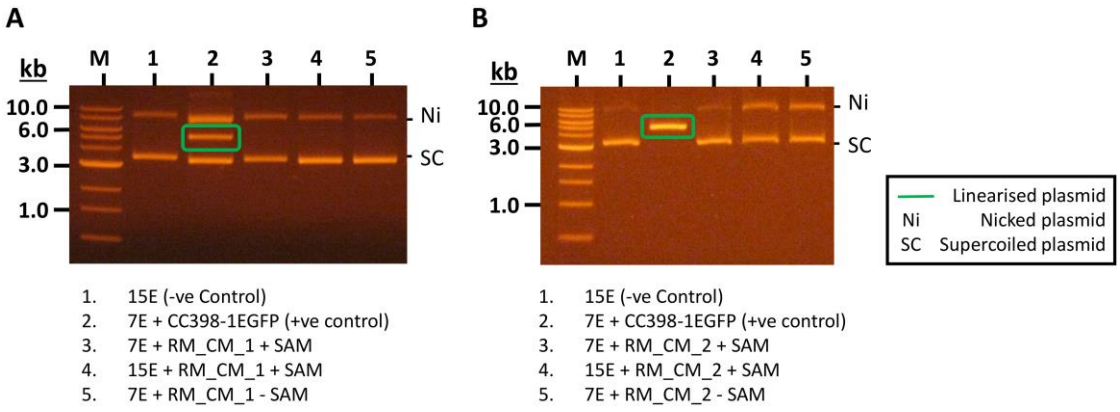


Figure 48: DNA cleavage assay using the RM_CM_1 protein (A) or the RM_CM_2 protein (B).

Results from the DNA cleavage assays showed no signs of restriction activity in either of the first two fusions and as such, new fusion proteins were designed and produced. The next two fusion genes to be expressed were RM_EB_1 and RM_EB_2 (2332 and 2337 bp fusion genes respectively). The genes were created by PCR, ligated into the pJF vector and expressed on a large scale. The subsequent recombinant proteins were then purified, samples from which were subjected to SDS-PAGE.

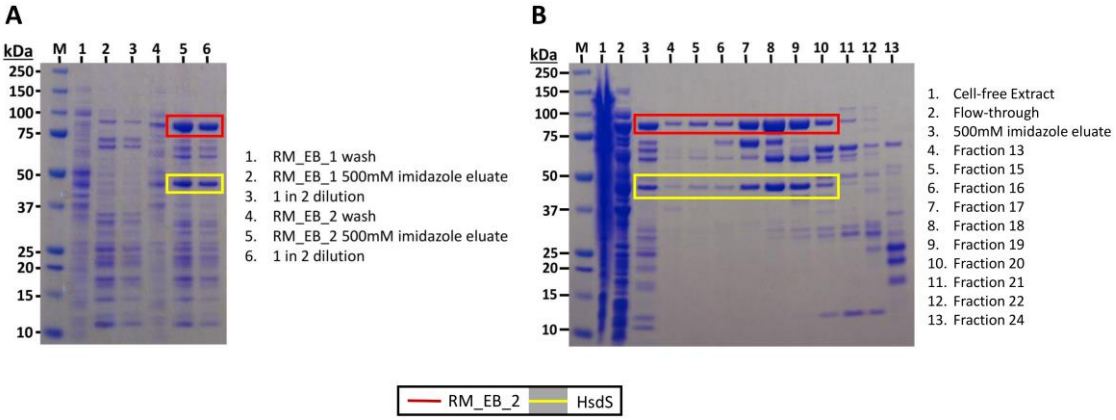


Figure 49: Nickel affinity purification of RM_EB_1 and 2 (A) and size exclusion purification of RM_EB_2 (B).

Results from SDS-PAGE analysis showed that RM_EB_2 had been successfully eluted from the nickel column, whereas RM_EB_1 had not (Fig. 49A). With an objective overview of these results, the reason for the absence of the RM_CM_1 and RM_EB_1 proteins after the nickel affinity step could be seen. The portion of the *hsdR* gene contained in the RM_CM_1 fusion is 758 bp. This means that transcription of the downstream *hsdM* gene would be out of frame by 1 bp, and subsequent translation produces several peptide fragments. In the case of the RM_EB_1 fusion, the *hsdR* is 775 bp in length, and so transcription of the *hsdM* is 2 bp out of frame. This too would not produce the desired protein. This also means that both proteins will not be produced with a C-terminal HisTag, and so cannot be purified via this nickel affinity chromatography purification method. For this reason, no further work was conducted with these two constructs. On the other hand, RM_EB_2 was purified by nickel affinity and so further purification by size exclusion was performed (Fig. 49B). The purified sample was then used in a cleavage assay to determine its restriction activity.

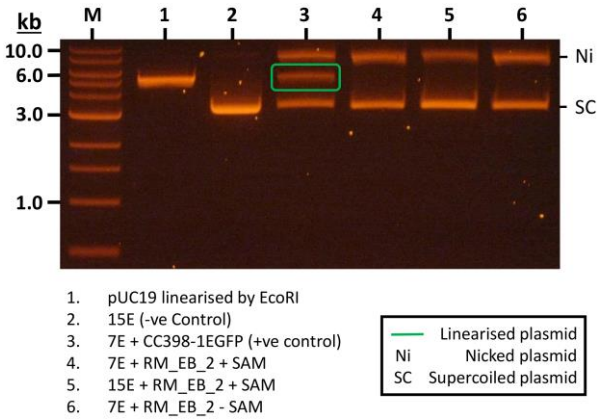


Figure 50: DNA cleavage assay using the RM_EB_2 protein.

The assay was conducted as before and returned no positive results (Fig. 50). It was thought that it was possible that the lack of identifiable activity was due to sub-optimal conditions, and so different conditions were tested. It had been previously discovered by the Dryden group, that the Sau1 R-M systems showed activity when the reaction was conducted in Buffer 4 (*New England Biolabs*). As such, this was the reaction buffer that was used for all subsequent reactions. However, Buffer 4 might not produce the ideal conditions for these novel fusion enzymes. Therefore, Buffers 1, 2 and 3 (*New England Biolabs*) were also tested in the plasmid cleavage assay for restriction activity in RM_EB_2. To rule out false positives due to contamination, each different buffer had ± SAM. The results of this experiment were run on an agarose gel and visualised on a transilluminator (Fig. 51).

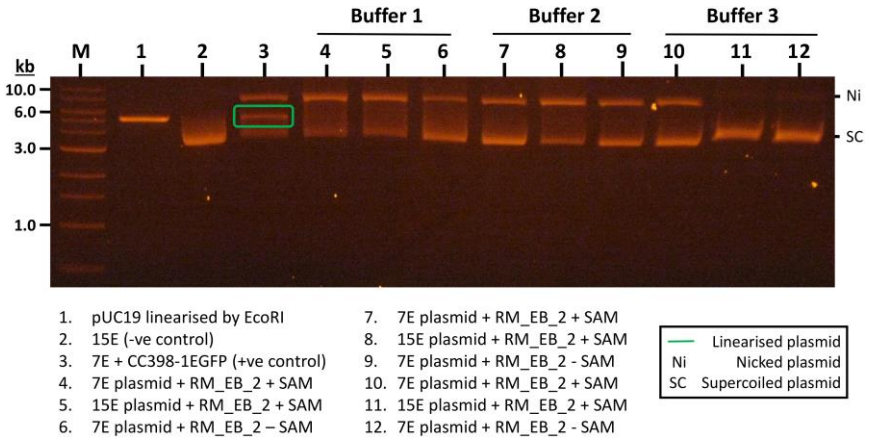


Figure 51: Plasmid cleavage assay for restriction activity on RM_EB_2, using Buffers 1, 2 and 3.

The multiple buffer assay proved inconclusive. Whilst Buffers 1 and 2 showed no discernible signs of a cleaved plasmid product, there was a difference using Buffer 3. Whilst the other test reactions contained a mixture of supercoiled and nicked species, Lanes 11 and 12 contained entirely supercoiled species. Although this occurred in two reactions that were expected to

give a negative result (due to lack of recognition sequence and lack of SAM respectively), this result was checked. A new plasmid cleavage assay was conducted, using only Buffer 3 (Buffer 4 for positive control). RM_EB_2 was incubated with "Eddys" 4E, 7E, 12E and 15E, with no SAM in two separate reactions containing two of the expected positives (7E and 12E) (Fig. 52).

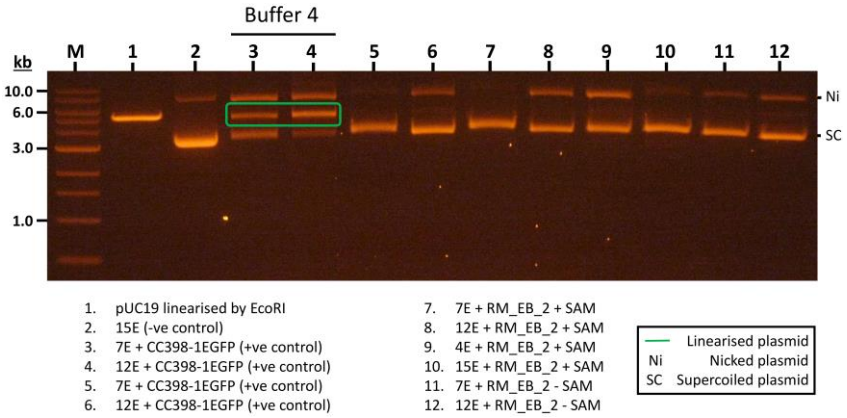


Figure 52: Plasmid cleavage assay for restriction activity in RM_EB_2, using Buffer 3.

The positive control reaction produced linearised plasmid species, whilst there seemed to be no complete cutting in the test reactions, and no identifiable pattern to nicked species. It was concluded from this that RM_EB_2 did not possess restriction activity.

The succession of negative results after a relatively large amount of preparation, suggested that a new method of investigating the novel R to M fusions would have to be found. The task would potentially involve high numbers of screening and as such, a high-throughput strategy was necessary. Success in *in vivo* assays of the engineered MTase proteins (shown later in this thesis), identified a possible method for assaying the R to M fusions. *In vivo* studies not only remove the need to produce a usable amount of soluble protein but also remove the concern over buffer conditions.

Nine more fusion genes (RM_EB_3 to 11) were designed and cloned. The PCR and cloning work for RM_EB_4 to 11 was performed by Dr John White (University of Edinburgh). The new genes, together with the previous two fusion genes (RM_CM_2 and RM_EB_2), were used to transform *E. coli* NM1261 cells, in preparation for the *in vivo* assays (Figs. 53 to 63).

N.B. The colour and contrast of images may have been adjusted, in order to present results more clearly.

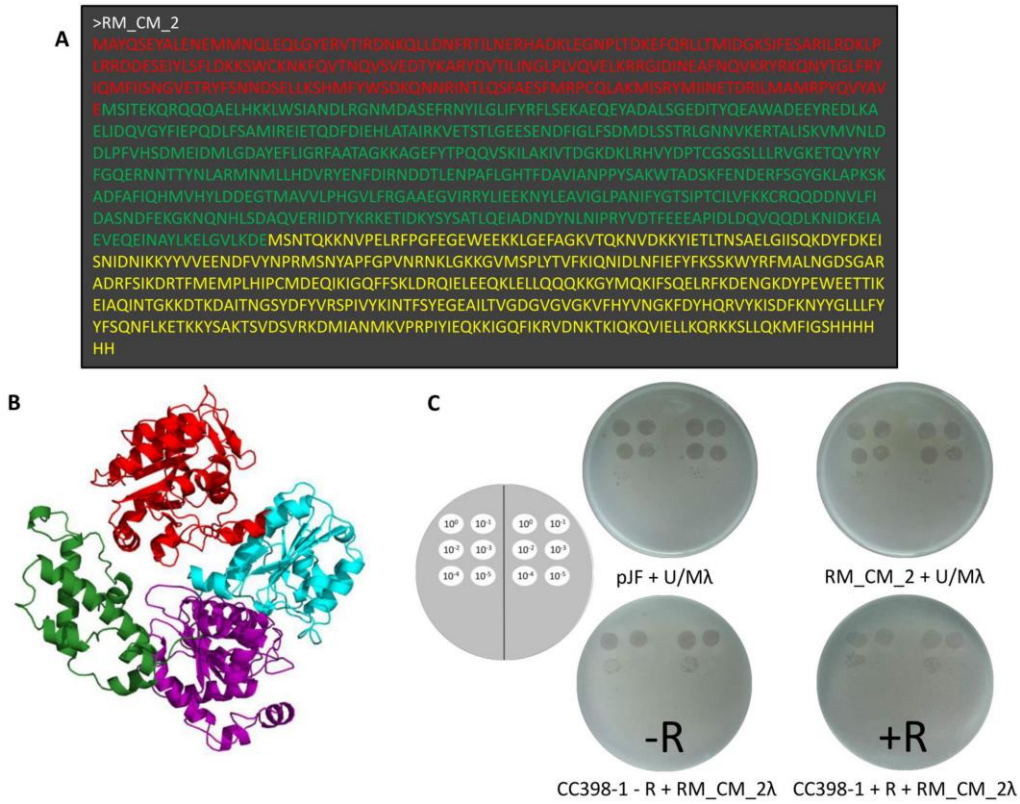


Figure 53: Amino acid sequence of RM_CM_2 (A), protein model of CC5 HsdR (B), and *in vivo* assay of RM_CM_2 (C). The section of the HsdR used in the fusion is highlighted in red. (A and B).

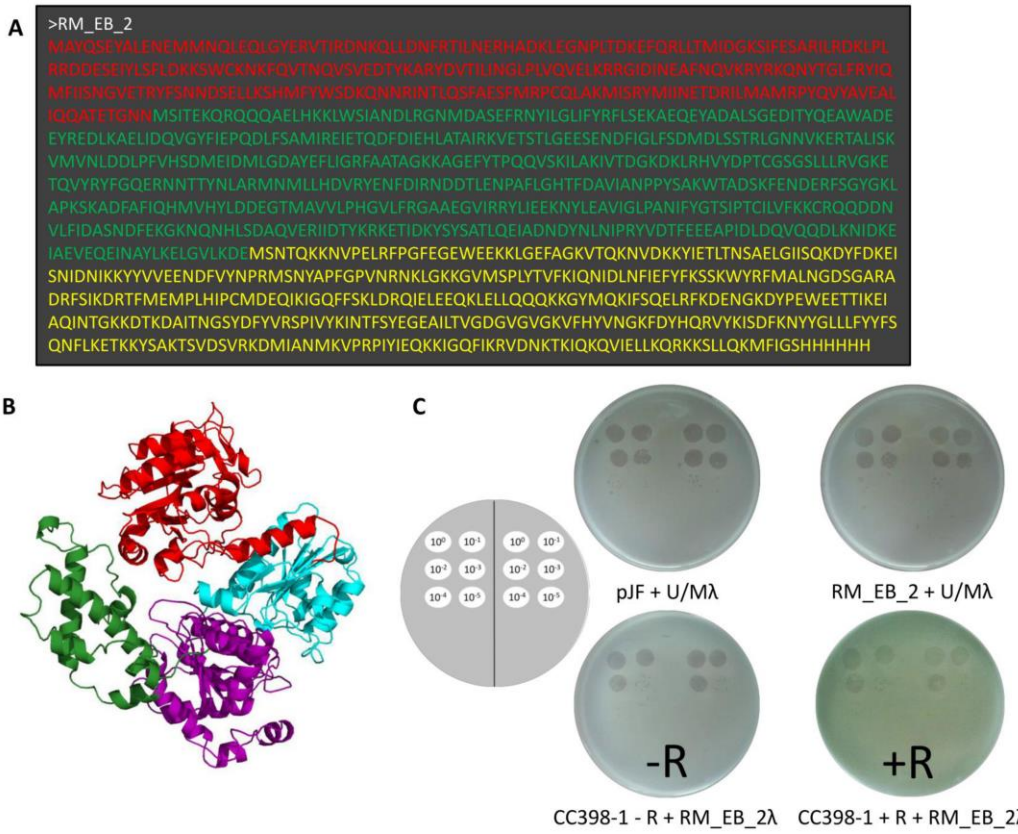


Figure 54: Amino acid sequence of RM_EB_2 (A), protein model of CC5 HsdR (B), and *in vivo* assay of RM_EB_2 (C). The section of the HsdR used in the fusion is highlighted in red. (A and B).

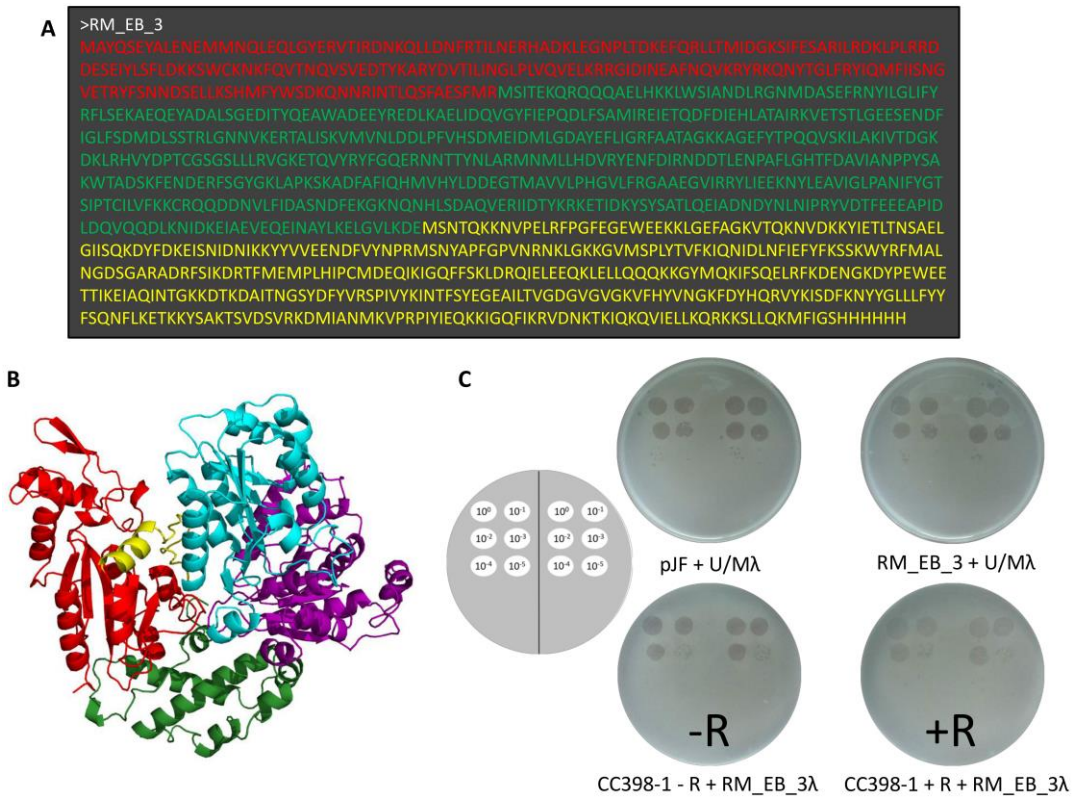


Figure 55: Amino acid sequence of RM_EB_3 (A), protein model of CC5 HsdR (B), and *in vivo* assay of RM_EB_3 (C). The section of the HsdR used in the fusion is highlighted in red. (A and B).

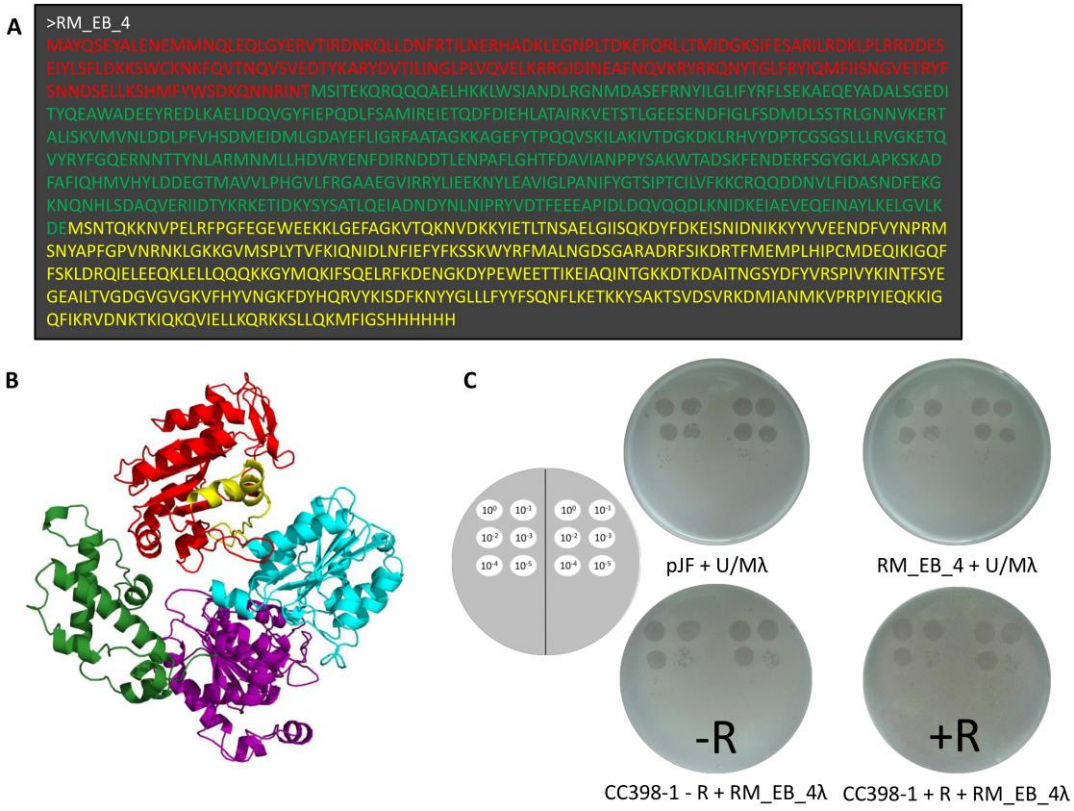


Figure 56: Amino acid sequence of RM_EB_4 (A), protein model of CC5 HsdR (B), and *in vivo* assay of RM_EB_4 (C). The section of the HsdR used in the fusion is highlighted in red. (A and B).

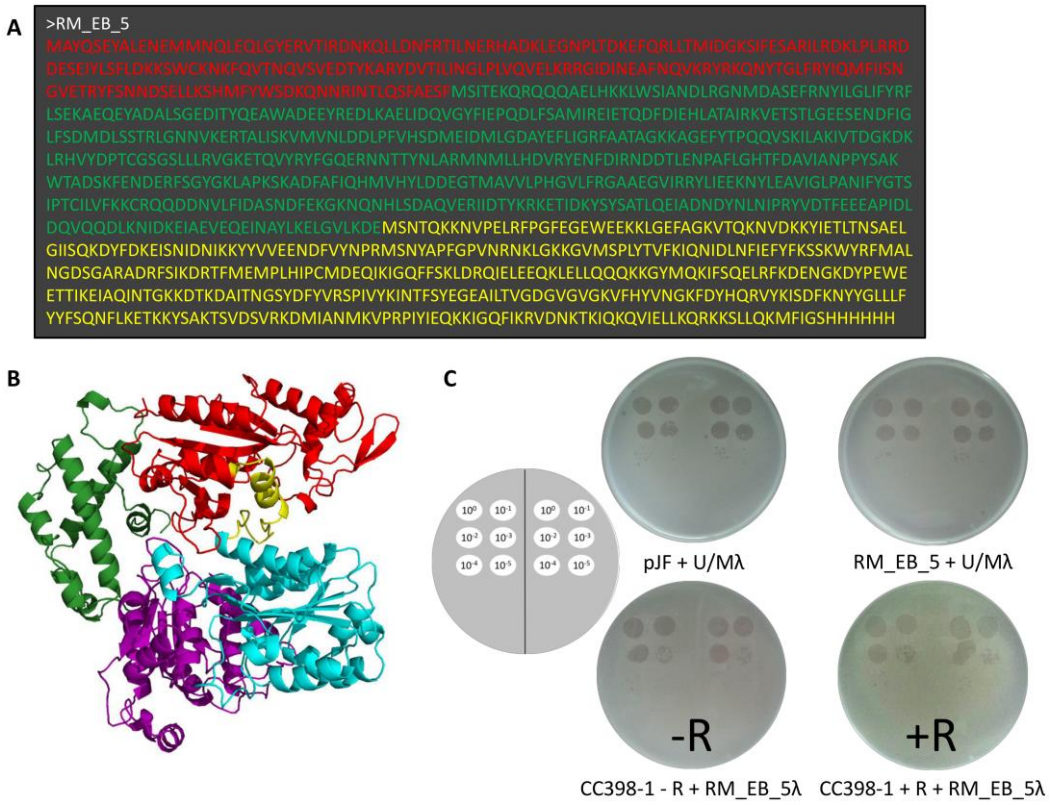


Figure 57: Amino acid sequence of RM_EB_5 (A), protein model of CC5 HsdR (B), and *in vivo* assay of RM_EB_5 (C). The section of the HsdR used in the fusion is highlighted in red. (A and B).

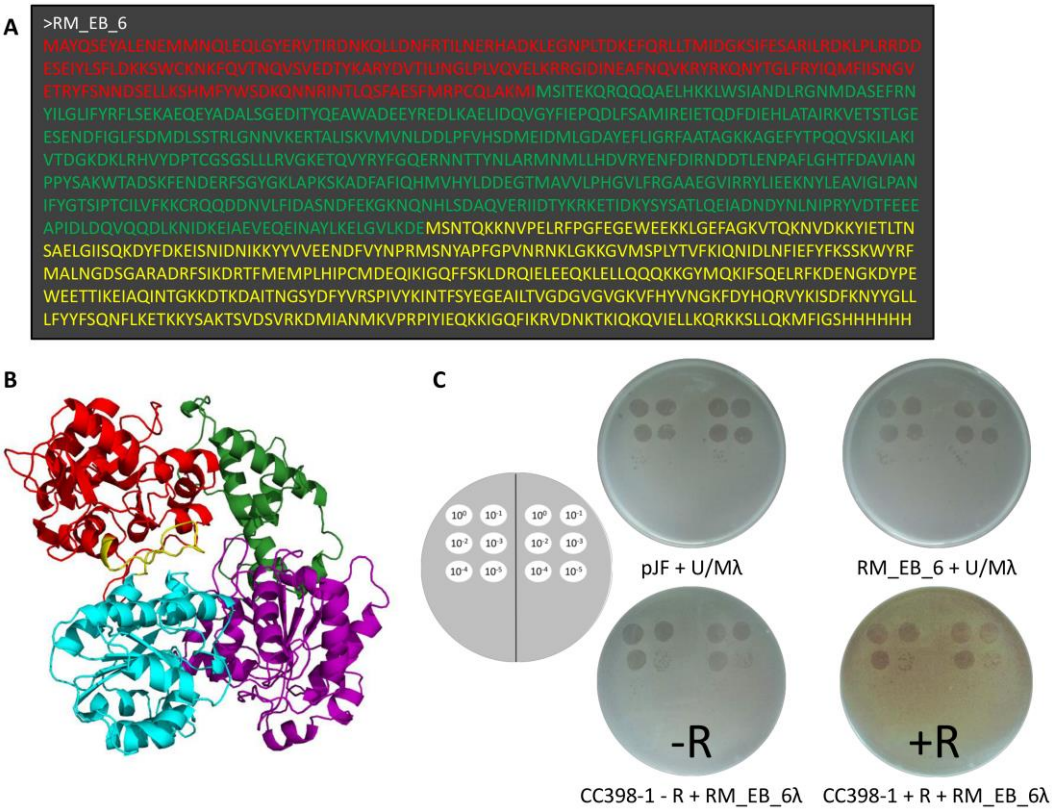


Figure 58: Amino acid sequence of RM_EB_6 (A), protein model of CC5 HsdR (B), and *in vivo* assay of RM_EB_6 (C). The section of the HsdR used in the fusion is highlighted in red. (A and B).

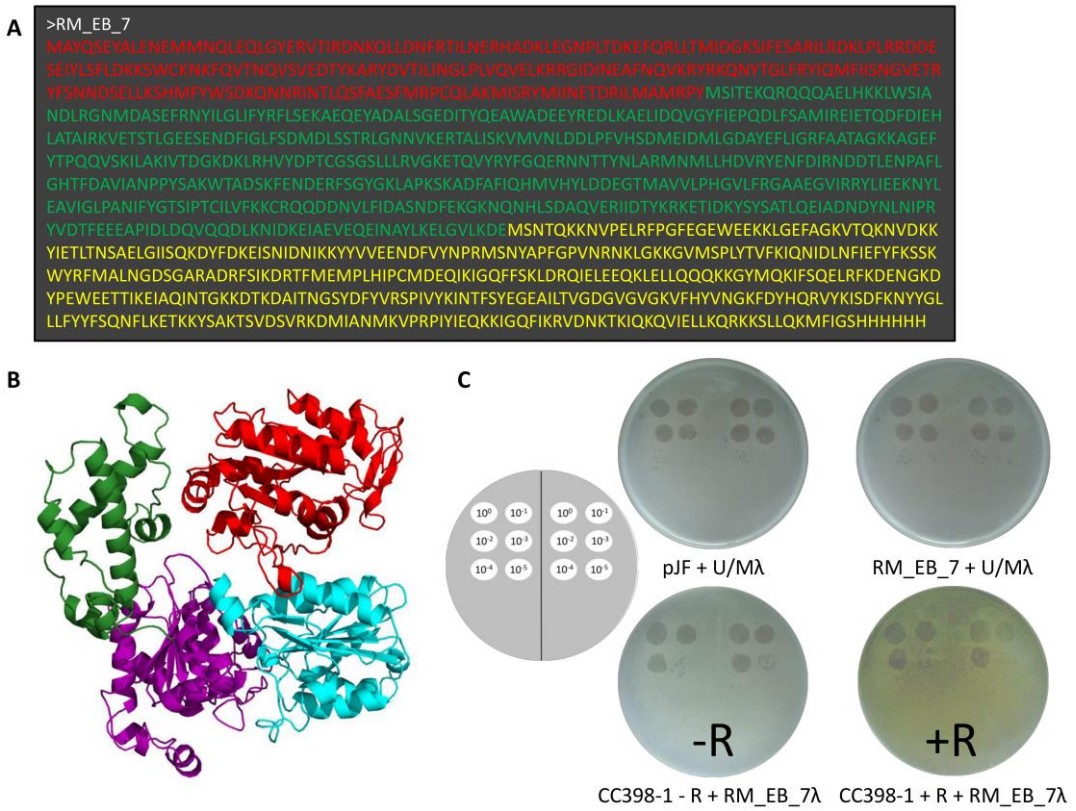


Figure 59: Amino acid sequence of RM_EB_7 (A), protein model of CC5 HsdR (B), and *in vivo* assay of RM_EB_7 (C). The section of the HsdR used in the fusion is highlighted in red. (A and B).

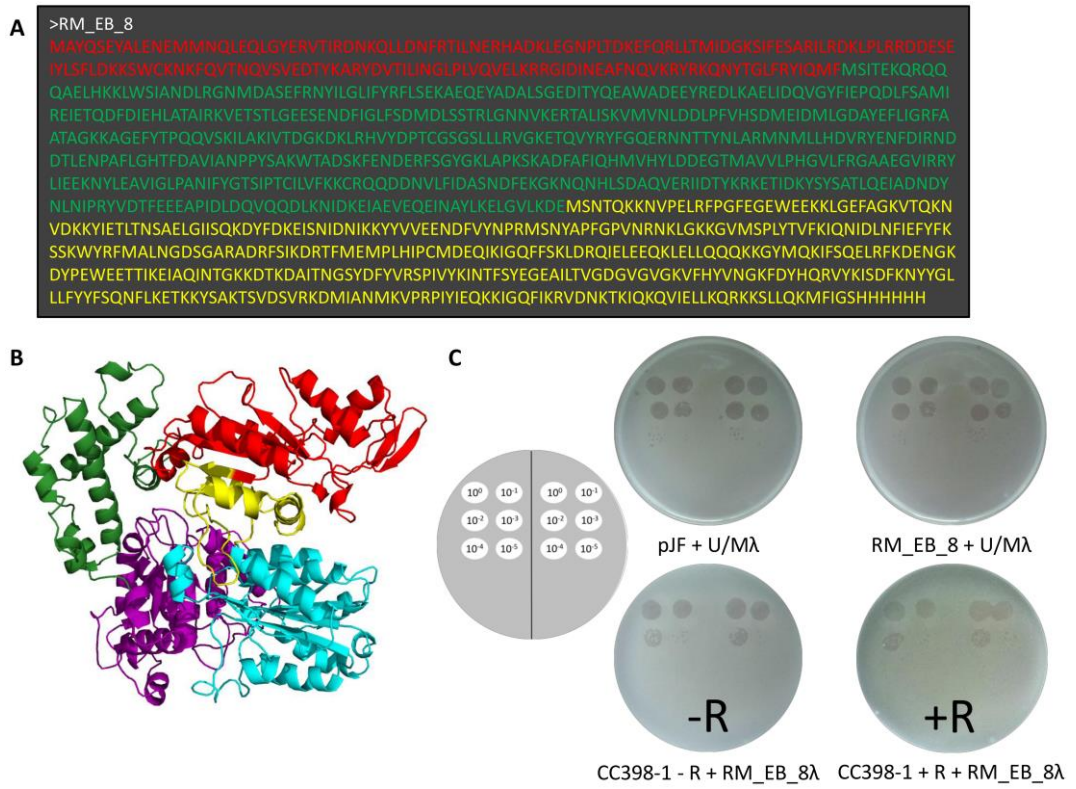


Figure 60: Amino acid sequence of RM_EB_8 (A), protein model of CC5 HsdR (B), and *in vivo* assay of RM_EB_8 (C). The section of the HsdR used in the fusion is highlighted in red. (A and B).

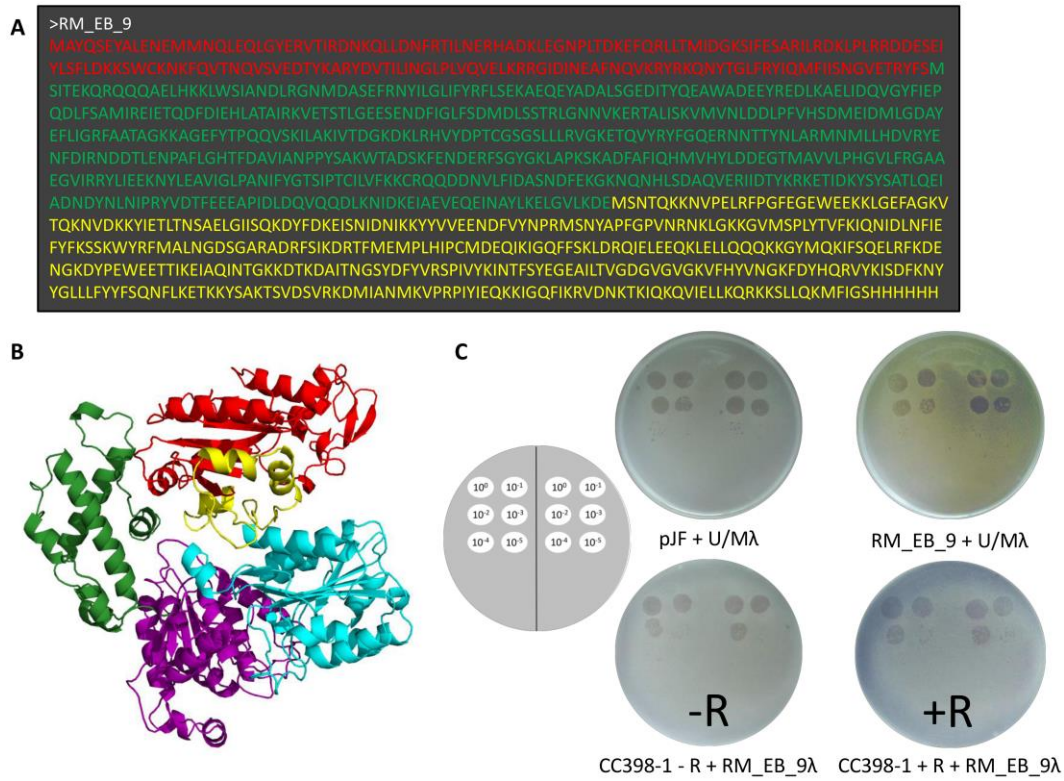


Figure 61: Amino acid sequence of RM_EB_9 (A), protein model of CC5 HsdR (B), and *in vivo* assay of RM_EB_9 (C). The section of the HsdR used in the fusion is highlighted in red. (A and B).

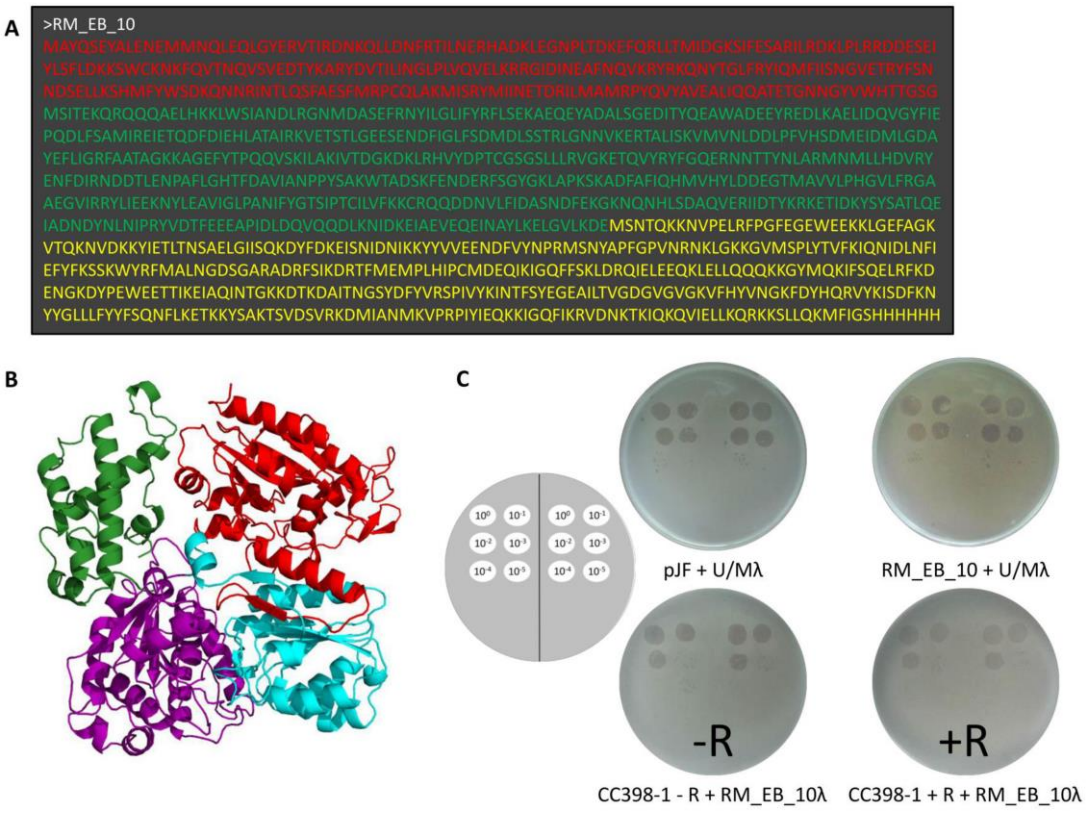


Figure 62: Amino acid sequence of RM_EB_10 (A), protein model of CC5 HsdR (B), and *in vivo* assay of RM_EB_10 (C). The section of the HsdR used in the fusion is highlighted in red. (A and B).

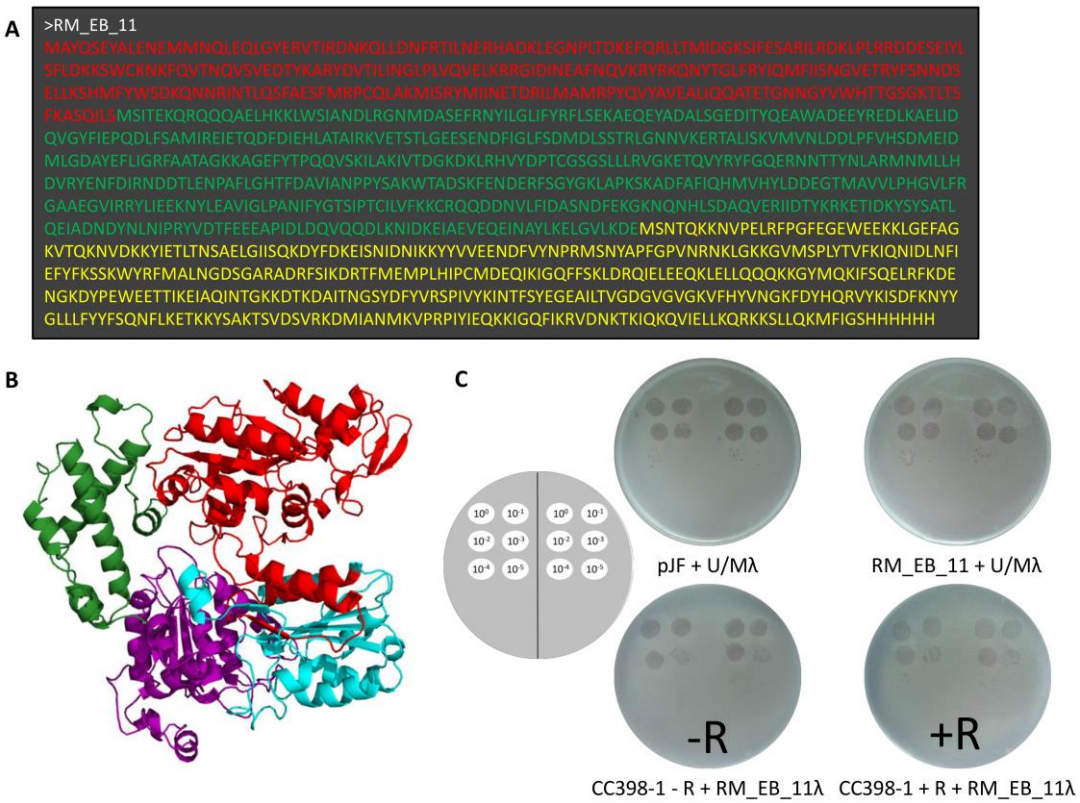


Figure 63: Amino acid sequence of RM_EB_11 (A), protein model of CC5 HsdR (B), and *in vivo* assay of RM_EB_11 (C). The section of the HsdR used in the fusion is highlighted in red. (A and B).

In vivo assays of the RM fusions yielded no signs of restriction activity. In every case, the dilution order of the phage plaques was the same in the control and RM fusion plates (Figs 53C to 63C).

As the RM fusions were created with an intact CC398-1 MTase, it was thought possible that they would retain their methylase activity. *E. coli* cells containing each of the RM fusions was infected with λ phage, which was then retrieved and used to infect *E. coli* cells containing the wild-type CC398-1 MTase genes and the CC5 *hsdR*. In every case, the dilution order of the phage plaques was the same in the plus and minus *hsdR* plates. This provided evidence that each of the RM fusions possessed an active MTase. This result is encouraging, as it suggests that the fusions are able to fold correctly and produce an active enzyme. Unfortunately, the sections of the HsdR that had been fused to this did not seem to be producing an active nuclease.

3.5. CC398-1 HsdM to HsdS Fusion

The two Type I MTase genes occur at the same locus but are expressed from separate frames. The first few base pairs at the 5' of the *hsdS* gene overlap with the 3' of the *hsdM* gene (Fig. 64). An MS fusion was created by the Dryden lab by removing the frameshift between the genes of the MTase from the EcoKI R-M system (Roberts et al. 2012). This resulted in joining the two MTase genes and producing a single peptide product. This construct showed R-M activities *in vivo* and restriction activity *in vitro*, when supplemented with stoichiometric amounts of free HsdM subunit. However, this had never been attempted with one of the Sau1 R-M systems.

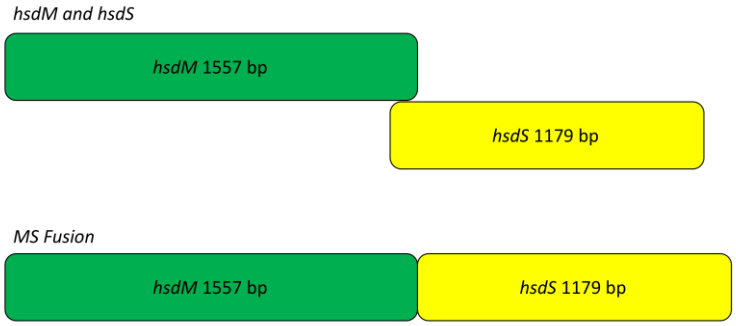
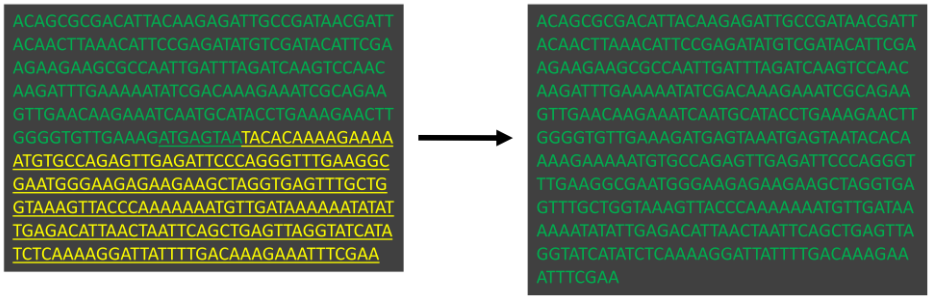


Figure 64: Cartoon diagram of the genetic arrangement of the wild-type (above) and MS fusion (below) MTases.

Primers to create an MS fusion gene from the CC398-1 MTase genes were designed and used in a PCR. The PCR was conducted in two stages and involved a set of primers complementary to both the 3' end of *hsdM* and the 5' of *hsdS* (see Materials and Methods for the Crossover PCR mechanism). After the second round of PCR, a single fusion sequence was created, which had removed the frameshift between the two genes (Fig. 65A). The CC398-1 MS fusion gene had been made successfully, and its nucleotide sequence was verified by Sanger sequencing.

The expression of the wild-type MTase genes results in two distinct subunits, which closely associate to form an active methyltransferase in an M_2S_1 conformation. Fusing these genes removes the gap between the subunits and results in an M_1S_1 fusion protein (Fig. 65B). Using the model of the CC398-1 MTase, this gap was estimated to be ~3.6 Angstroms, a distance that is eliminated in the fusion protein (Fig. 65D).

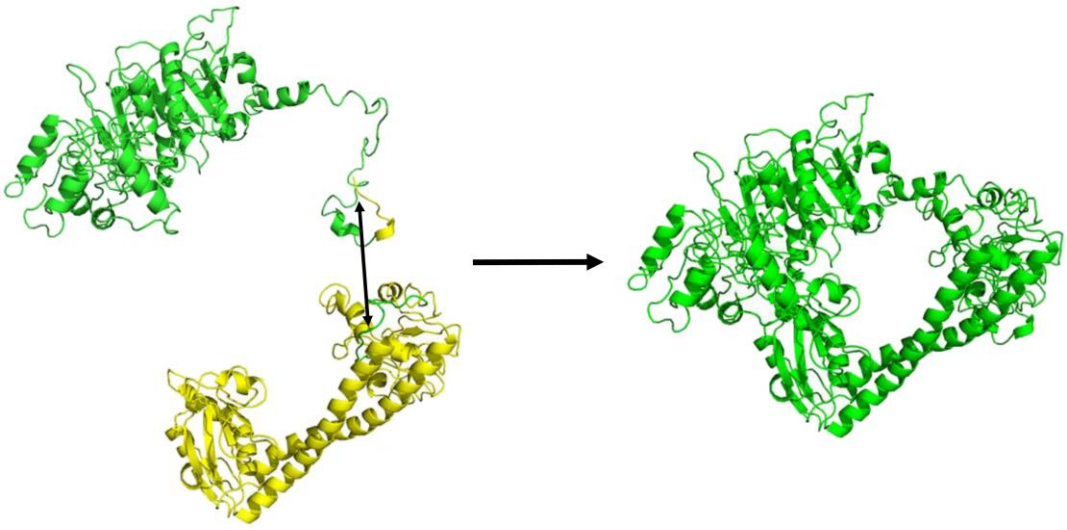
A


ACAGCGCGACATTACAAGAGATTGCCGATAACGATT
ACAACTTAAACATTCCGAGATATGTCGATACATTCGA
AGAAGAAGCGCCAATTGATTTAGATCAAGTCCAAC
AAGATTTGAAAAATATCGACAAAGAAATCGCAGAA
GTTGAACAAGAAATCAATGCATACCTGAAAGAACTT
GGGGTGTTGAAAGATGAGTAATACACAAAAGAAAA
ATGTGCCAGAGTTGAGATTCCCAGGGTTTGAAGGC
GAATGGGAAGAGAAGAAGCTAGGTGAGTTTGCTG
GTAAAGTTACCCAAAAAAATGTTGATAAAAAATATAT
TGAGACATTAACTAATTCAGCTGAGTTAGGTATCATA
TCTCAAAAGGATTATTTTGACAAAGAAATTTCGAA
ACAGCGCGACATTACAAGAGATTGCCGATAACGATT
ACAACTTAAACATTCCGAGATATGTCGATACATTCGA
AGAAGAAGCGCCAATTGATTTAGATCAAGTCCAAC
AAGATTTGAAAAATATCGACAAAGAAATCGCAGAA
GTTGAACAAGAAATCAATGCATACCTGAAAGAACTT
GGGGTGTTGAAAGATGAGTAAATGAGTAATACACA
AAAGAAAAATGTGCCAGAGTTGAGATTCCCAGGGT
TTGAAGGCGAATGGGAAGAGAAGAAGCTAGGTGA
GTTTGCTGGTAAAGTTACCCAAAAAAATGTTGATAA
AAAATATATTGAGACATTAACTAATTCAGCTGAGTTA
GGTATCATATCTCAAAAGGATTATTTTGACAAAGAA
ATTTCGAA

B


MSITEKQRQQQAELHKKLWSIANDLRGNMDASEFRNYILGLIF
YRFLSEKAEQEYADALSGEDITYQEAWADEEYREDLKAELIDQV
GYFIEPQDLFSAMIREIETQDFDIEHLATAIRKVETSTLGEESEND
FIGLFSDMDLSSTRLGNNVKERTALISKVMVNLDDLPFVHSDM
EIDMLGDAYEFLIGRFAATAGKKAGEFYTPQQVSKILAKIVTDG
KDKLRHVYDPTCGSGSLLLRVGKETQVYRYFGQERNNTTYNLA
RMNMLLHDVRYENFDIRNDDTLENPAFLGHTFDAVIANPPYS
AKWTADSKFENDERFSGYGKLAPKSKADFAFIQHMVHYLDDE
GTMAVVLPHGVLFRGAAEGVIRRYLIEEKNYLEAVIGLPANIFY
GTSIPTCILVFKKCRQQDDNVLFIDASNDFEKGKNQNHLSDAQ
VERIIDTYKRKETIDKYSYSATLQEIADNDYNLNIPRYVDTFEEEA
PIDLDQVQQDLKNIDKEIAEVEQEINAYLKELGVLKDEMSNTQK
KNVPELRFPGFEGEWEEKKLGEFAGKVTQKNVDKKYIETLTNSA
ELGIISQKDYFDKEISNIDNIKKYYVVEENDFVYNPRMSNYAPF
GPVNRNKLGKKGVMSPLYTVFKIQNIDLNFIEFYFKSSKWYRF
MALNGDSGARADRFSIKDRTFMEMPLHIPCMDEQIKIGQFFS
KLDRQIELEEQKLELLQQQKKGYMQKIFSQELRFKDENGKDYPE
WEETTIKEIAQINTGKKDTKDAITNGSYDFYVRSPIVYKINTFSYE
GEAILTVGDGVGVGKVFHYVNGKFDYHQRVYKISDFKNYYGLL
LFYYFSQNFLKETKKYSAKTSVDSVRKDMIANMKVPRPIYIEQK
KIGQFIKRVDNKTKIQKQVIELLKQRKKSLLQKMFIGSHHHHHH
MSITEKQRQQQAELHKKLWSIANDLRGNMDASEFRNYILGLIF
YRFLSEKAEQEYADALSGEDITYQEAWADEEYREDLKAELIDQV
GYFIEPQDLFSAMIREIETQDFDIEHLATAIRKVETSTLGEESEND
FIGLFSDMDLSSTRLGNNVKERTALISKVMVNLDDLPFVHSDM
EIDMLGDAYEFLIGRFAATAGKKAGEFYTPQQVSKILAKIVTDG
KDKLRHVYDPTCGSGSLLLRVGKETQVYRYFGQERNNTTYNLA
RMNMLLHDVRYENFDIRNDDTLENPAFLGHTFDAVIANPPYS
AKWTADSKFENDERFSGYGKLAPKSKADFAFIQHMVHYLDDE
GTMAVVLPHGVLFRGAAEGVIRRYLIEEKNYLEAVIGLPANIFY
GTSIPTCILVFKKCRQQDDNVLFIDASNDFEKGKNQNHLSDAQ
VERIIDTYKRKETIDKYSYSATLQEIADNDYNLNIPRYVDTFEEEA
PIDLDQVQQDLKNIDKEIAEVEQEINAYLKELGVLKDEMSNTQK
KNVPELRFPGFEGEWEEKKLGEFAGKVTQKNVDKKYIETLTNSA
ELGIISQKDYFDKEISNIDNIKKYYVVEENDFVYNPRMSNYAPF
GPVNRNKLGKKGVMSPLYTVFKIQNIDLNFIEFYFKSSKWYRF
MALNGDSGARADRFSIKDRTFMEMPLHIPCMDEQIKIGQFFS
KLDRQIELEEQKLELLQQQKKGYMQKIFSQELRFKDENGKDYPE
WEETTIKEIAQINTGKKDTKDAITNGSYDFYVRSPIVYKINTFSYE
GEAILTVGDGVGVGKVFHYVNGKFDYHQRVYKISDFKNYYGLL
LFYYFSQNFLKETKKYSAKTSVDSVRKDMIANMKVPRPIYIEQK
KIGQFIKRVDNKTKIQKQVIELLKQRKKSLLQKMFIGSHHHHHH

C



D

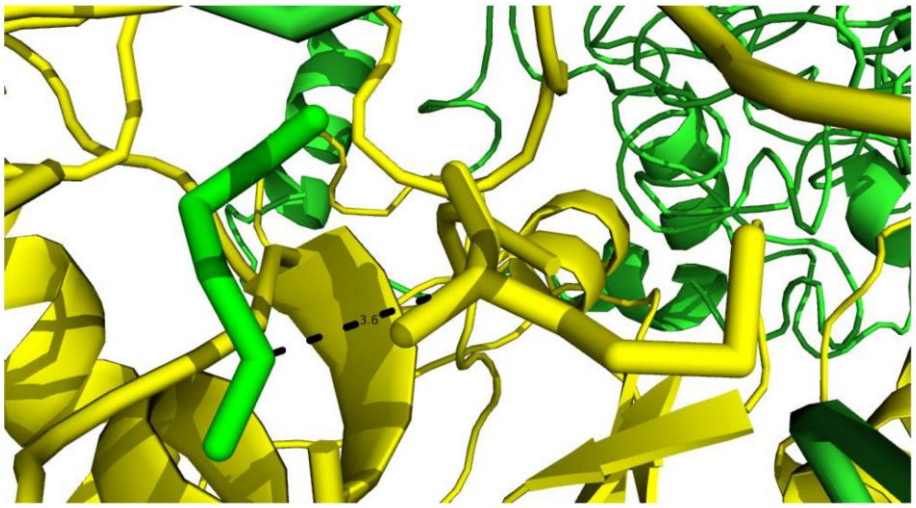


Figure 65: Sections of CC398-1 MTase and MS fusion nucleotide sequences, showing the removal of the overlap between the *hsdM* and *hsdS* genes (A). Amino acid sequence of the two CC398-1 MTase subunits becoming one peptide (B). Cartoon of the subunits becoming one single peptide (C). The C-terminal of HsdM is coloured yellow and the N-terminal of HsdS is coloured green to show where they will be joined. The CC398-1 subunit interface (D). The C-terminal glutamate of the HsdM and the N-terminal methionine of the HsdS. In this model of the CC398-1 MTase, the α-carbons in these residues are 3.6 Angstroms apart.

The fusion gene was successfully ligated into vector pJF (see Appendix C for plasmid map), and was cloned by using the product to transform *E. coli* DH5-α cells. After the successful cloning of the MS fusion gene in vector pJF (subsequently known as pJFMSF), it was then used in expression studies. Expression of the MS fusion gene was initially conducted on a small scale. It was induced with IPTG and left for ~3 hours at 37 ^{o}C. The cells were then pelleted, resuspended in buffer, lysed by sonication and then centrifuged to pellet the insoluble fraction. The results of this separation were analysed by SDS-PAGE (Fig. 66).

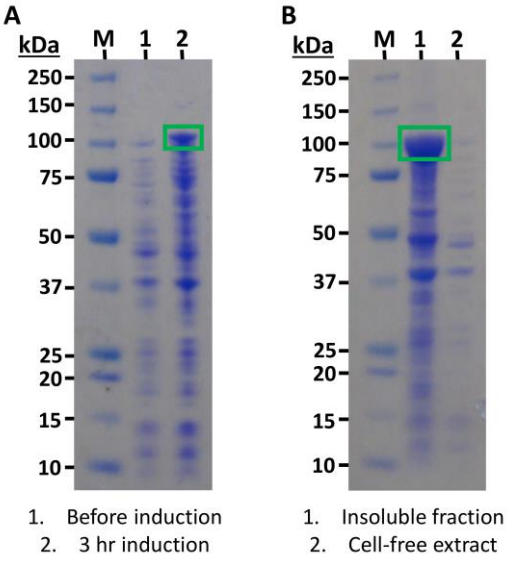


Figure 66: SDS-PAGE analysis of the first inductions of the CC398-1 MS fusion gene (A). To test the solubility of the fusion protein, the cells were suspended in buffer, lysed and centrifuged, separating the mixture into two fractions. These were analysed by SDS-PAGE (B). The green box highlights an over-expressed protein at the expected molecular weight of the fusion protein.

SDS-PAGE showed that the MS fusion gene had expressed but that none had been recovered in the cell-free extract. Under these conditions, the MS fusion protein was insoluble, and so a new method of investigation had to be employed. The *in vivo* assays proved a relatively quick and convenient way of identifying activity, and so this method was also used for the MS fusion (Fig. 67).

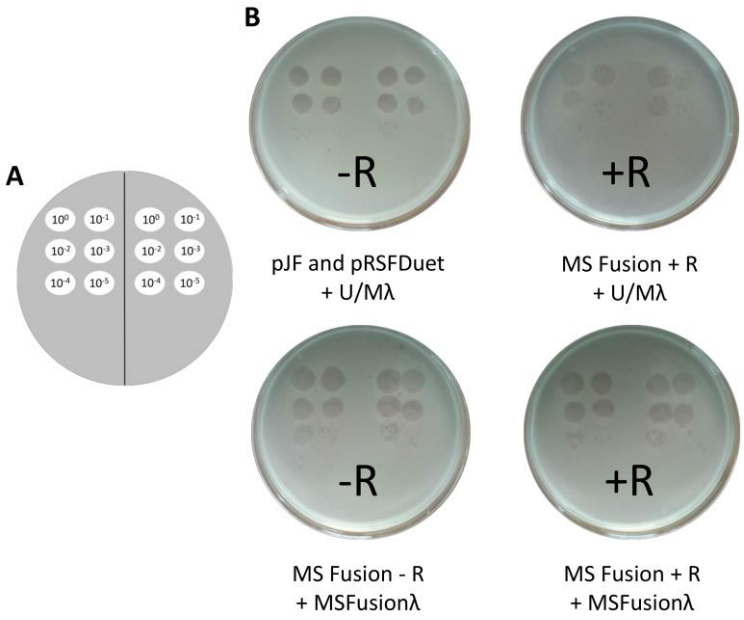


Figure 67: Diagram of spot test dilutions (A) *in vivo* spot test assay of the CC398-1 MS fusion (B).

Preliminary *in vivo* spot tests of the MS fusion showed evidence of both restriction and modification activities. The plaques formed by the MS fusion modified λ appeared to be the same in both the + and – R plates, whilst there appeared to be a significant level of cut back to the unmodified (U/M) λ (at least one log difference). To verify this result, full plate assays were conducted (Table 3).

MS fusion (MSF) MTase *In vivo* Assay Results:

Phage Type	Phage Dilution	Phage Volume	R-M System	Number of Plaques
Unmodifiedλ	10^{-5}	22.5 µL	None	384
	10^{-4}	35 µL	MSF + R	650
	10^{-6}	100 µL	None	192
	10^{-5}	223 µL	MSF + R	808
	10^{-6}	100 µL	None	274
	10^{-5}	40 µL	MSF + R	34
	10^{-5}	100 µL	None	1516
	10^{-6}	40 µL	MSF + R	6
	10^{-6}	100 µL	None	263
	10^{-5}	40 µL	MSF + R	66
MSFλ	10^{-5}	32.3 µL	MSF	401
	10^{-5}	38.5 µL	MSF + R	511
	10^{-6}	167 µL	MSF	219
	10^{-6}	400 µL	MSF + R	383
	10^{-6}	200 µL	MSF	368
	10^{-6}	200 µL	MSF + R	241
	10^{-6}	200 µL	MSF	341
	10^{-6}	200 µL	MSF + R	237
	10^{-6}	200 µL	MSF	292
	10^{-6}	200 µL	MSF + R	247

Table 3: Raw data from several rounds of full plate *in vivo* assays of the MS fusion MTase. The table shows ±R pairs, which are the results from the experiment (+R) and control (-R), and the subsequent repeats. The plaque numbers cannot be compared without adjusting for volume and dilution.

The E.O.P. for the unmodified λ phage against the restriction active MS fusion MTase was 0.09 ±0.05 (see Appendix E for calculations). Comparing this to the wild-type value (0.15 ±0.11), indicated that the restriction active MS fusion was cutting back the phage to a similar extent as the wild-type. This value showed that the restriction complex is active. The E.O.P. of the MS fusion (MSF) modified λ phage was 0.78 ±0.14. Given the value calculated for the wild-type enzyme was 0.88 ±0.27, this suggests that the MS fusion is a comparable MTase. However, the high degree of uncertainty inherent in this assay means that these results were considered to be within experimental error from those gathered from the wild-type enzyme.

As such, no further assumptions could be made. Nevertheless, they do confirm that the MS fusion forms an active R-M system.

In vivo analysis had provided evidence that the MS fusion protein could modify λ phage DNA. It was assumed that it could also modify the genomic DNA in *E. coli* cells and as such, could be used in the process of SMRT sequencing. The pJFMSF plasmid was used to transform *E. coli* ER2796 cells, from which an overnight culture was established. The genomic DNA was purified from these cells and then screened by SMRT sequencing (*Pacific Biosciences*).

MS fusion (MSF) MTase SMRT Sequencing Results:

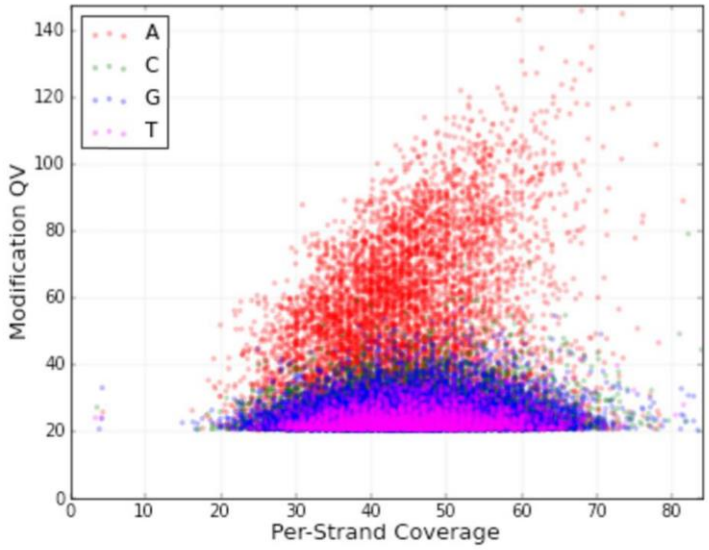
A

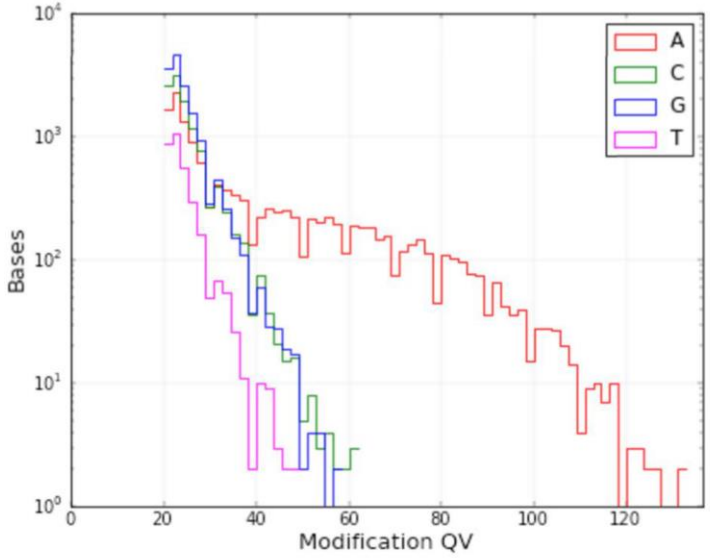
Reports for Job Bower_MSF_10915

PACIFIC BIOSCIENCES

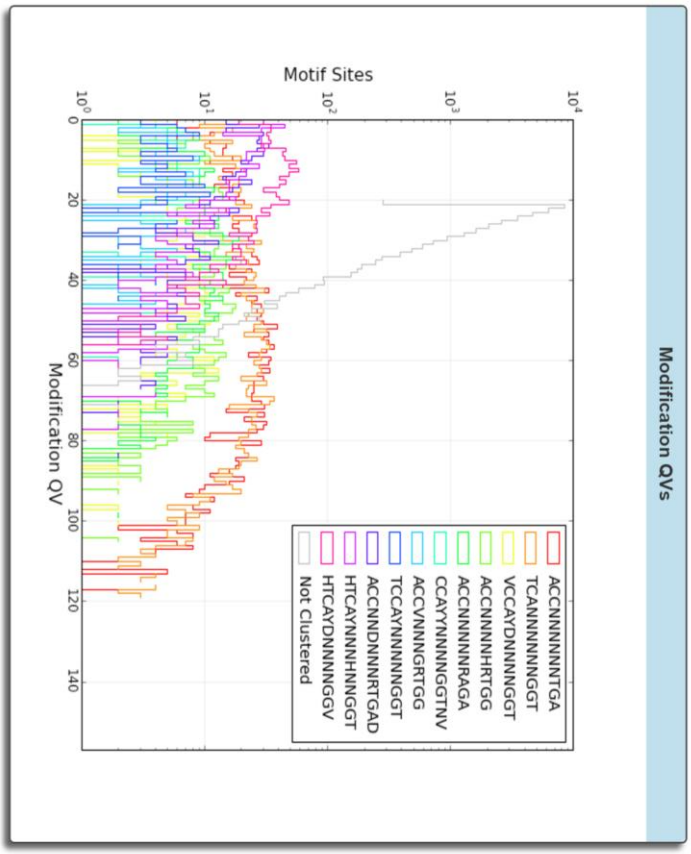
SMRT Cells: 2 Movies: 2

Kinetic Detections

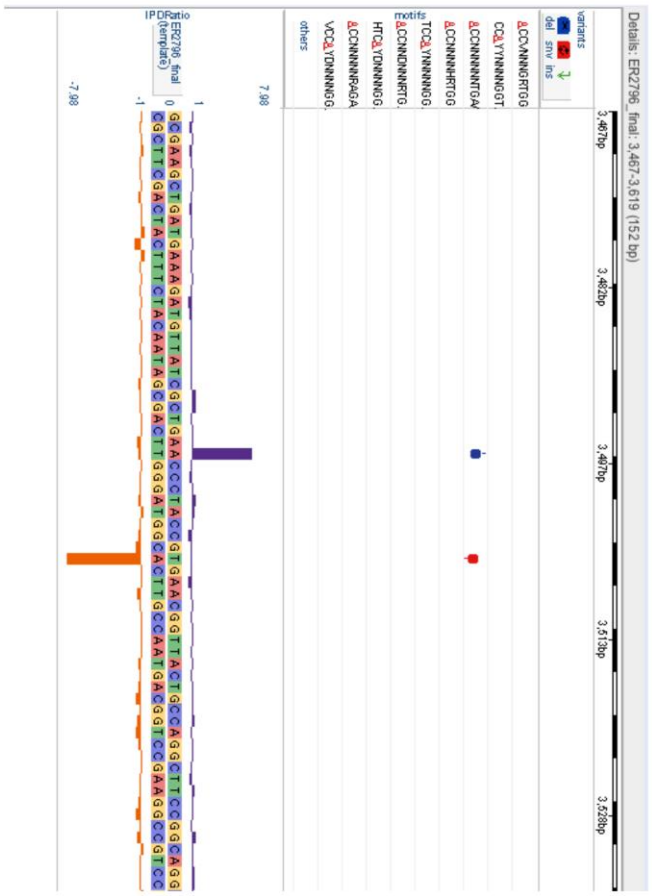




B


Modification QVs
Motif Sites
Modification QV
ACCNNNNNNTGA
TCANNNNNNGGT
VCCAYDNNNNGGT
ACCNNNNHRTGG
ACCNNNNNRAGA
CCAYYNNNNGGTNV
ACCVNNNGRTGG
TCCAYNNNNNGGT
ACCNNDNNNRTGAD
HTCAYNNNHNNGGT
HTCAYDNNNNGGV
Not Clustered

C


Details: ER2796_final: 3,467-3,619 (152 bp)
variants
del
snv
ins
motifs
others
IPDRatio
ER2796_final
(template)

D

Reports for Job Bower_MSF_10915

PACIFIC BIOSCIENCES

SMRT Cells: 2 Movies: 2

Motif Summary

Motifs	Modified Position	Type	% Motifs Detected	# Of Motifs Detected	# Of Motifs In Genome	Mean Modification QV	Mean Motif Coverage	Partner Motif
ACCNNNNNNTGA	1	m6A	76.88%	1663	2163	63.45	43.68	TCANNNNNNGGT
TCANNNNNNGGT	3	m6A	76.05%	1645	2163	64.29	43.92	ACCNNNNNNTGA
VCCAYDNNNNGGT	4	m6A	76.05%	451	593	56.52	44.72	
ACCNNNNHRTGG	1	m6A	69.46%	523	753	54.15	43.83	
ACCNNNNNRAGA	1	m6A	58.35%	388	665	50.50	44.44	
CCAYYNNNNGGTNV	3	m6A	57.68%	244	423	56.54	43.71	
ACCVNNNGRTGG	1	m6A	38.86%	82	211	48.91	43.68	
TCCAYNNNNNGGT	4	m6A	29.36%	64	218	46.92	42.88	
ACCNNDNNNRTGAD	1	m6A	24.23%	181	747	47.75	45.98	HTCAYNNNHNNGGT
HTCAYNNNHNNGGT	4	m6A	22.49%	168	747	47.97	46.16	ACCNNDNNNRTGAD
HTCAYDNNNNGGV	4	m6A	18.75%	261	1392	40.10	46.18	

Figure 68: Results from SMRT sequencing (*Pacific Biosciences*) of MS fusion modified *E. coli* genomic DNA. Kinetic Detections (A). The first graph is a scatter plot showing the detection of base methylation. The higher coverage and detection of adenine methylation (red dots) confirms the MS fusion is an adenine methylase. The second graph shows detection of bases against detection of methylated bases. This too confirms adenine methylation. DNA sequence motif detection (B). The graph shows DNA sequence motif detection against detection of methylated motifs. $ACCN_6TGA$ (Red) and $TCAN_6GGT$ (Orange) are the most frequently detected motifs. Example of a detected motif (C). The bars are proportional to the time taken for the nucleotide to be incorporated into the growing chain. The high bars on the top and bottom adenines indicate that it was a relatively longer period to add these bases, and they are therefore recognised as being methylated. Table to show statistical data for the modified motifs (D).

As the MS fusion had been created without any alteration to the HsdS subunit, it was not expected to have a modification activity different from that of the wild-type CC398-1 enzyme. Results from SMRT sequencing confirmed this assumption was correct, but also raised some questions (Fig. 68). As expected, the MS fusion strongly preferred to methylate adenine, and had modified over 76 % of the wild-type DNA recognition sequences (ACCNNNNNRTGA) occurring in the *E. coli* genome. The second highest occurring sequence (also at over 76 %) was the reverse complement of the target sequence. However, analysis of the sequencing data by the *Pacific Biosciences* software not only indicated that the enzyme was methylating at other sites, but also that it was not necessarily a purine (R) at the 9th position of the sequence (Fig. 68D). Further scrutiny of the results suggests that some of the motifs generated may have been called erroneously. For example, the 5th motif from the top of the list, ACCNNNNNRAGA, would not be possible, as it suggests that the MTase is methylating a thymine on the bottom strand.

To clarify these results, further data analysis was conducted using *New England Biolabs* software. The R at position 9 of the recognition sequence was substituted with each base (ie. ACCNNNNN(A/T/C/G)TGA), and checked against the collected data for percentage occurrence (Fig. 69).

PacBio motif stats

Job ID: 21717
Specificity: ACCNNNNNGTGA
IPD threshold: 2
Modifications source: (•) GFF (fast) ( ) CSV (slow)
Calculate

Finding ACCNNNNNGTGA sites in ER2796
509 sites found
Loading modifications.gff
Processing ER2796
Total sites: 509

% of sites with IPD >= 2:	100.0	1.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0
average IPD:	5.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	A	**C**	**C**	**N**	**N**	**N**	**N**	**N**	**G**	**T**	**G**	**A**
	T	**G**	**G**	**N**	**N**	**N**	**N**	**N**	**C**	**A**	**C**	**T**
average IPD:	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.7	0.0	0.0
% of sites with IPD >= 2:	0.0	0.0	0.6	1.0	0.2	0.0	0.2	0.2	0.0	100.0	0.0	0.0

E-mail: vincze@neb.com

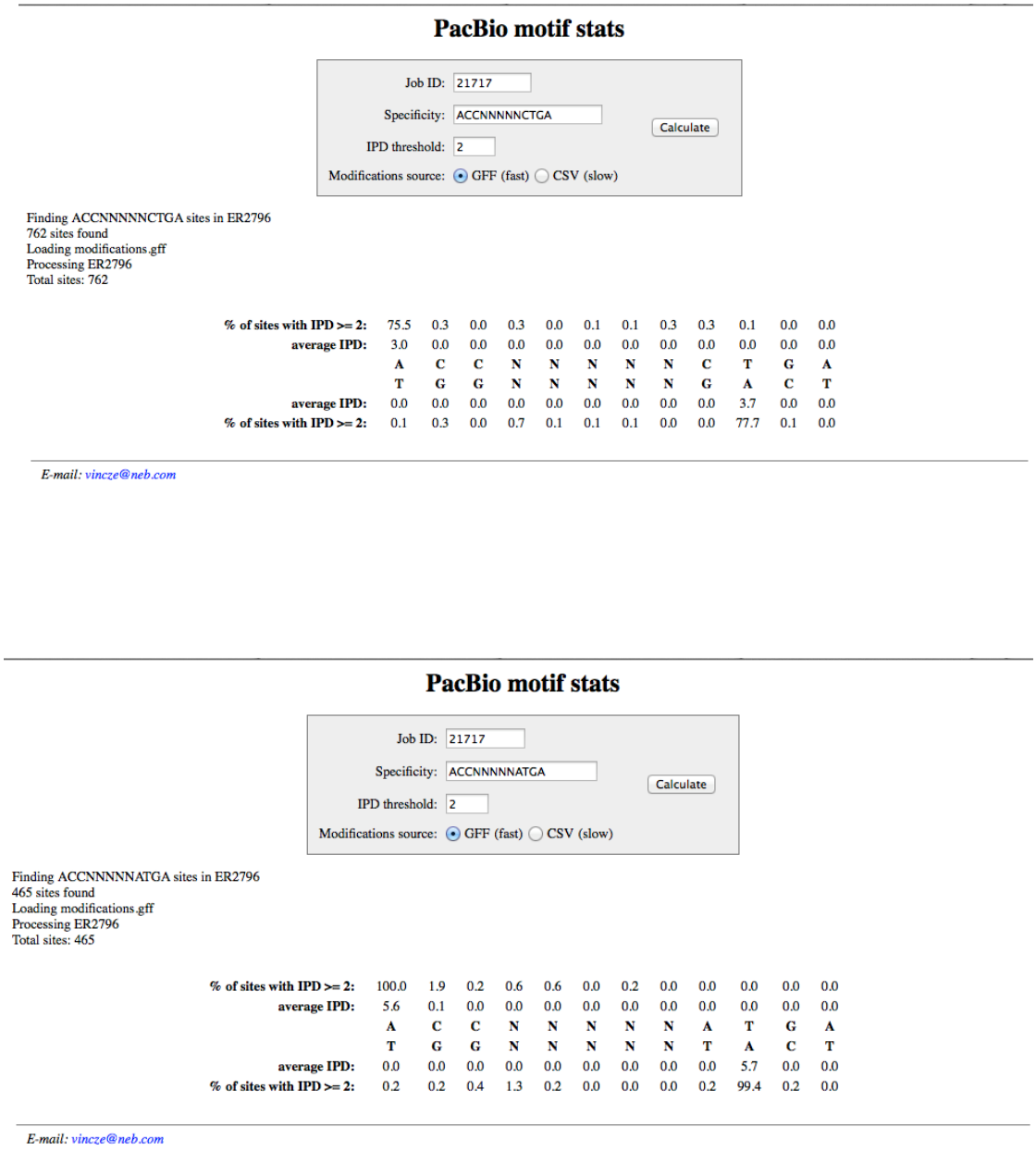
PacBio motif stats

Job ID: 21717
Specificity: ACCNNNNNTTGA
IPD threshold: 2
Modifications source: (•) GFF (fast) ( ) CSV (slow)
Calculate

Finding ACCNNNNNTTGA sites in ER2796
427 sites found
Loading modifications.gff
Processing ER2796
Total sites: 427

% of sites with IPD >= 2:	70.0	0.0	0.0	0.2	0.2	0.0	0.2	0.2	0.0	0.0	0.0	0.0
average IPD:	2.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	A	**C**	**C**	**N**	**N**	**N**	**N**	**N**	**T**	**T**	**G**	**A**
	T	**G**	**G**	**N**	**N**	**N**	**N**	**N**	**A**	**A**	**C**	**T**
average IPD:	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.2	0.0	0.0
% of sites with IPD >= 2:	0.0	0.0	0.5	1.2	0.0	0.0	0.5	0.0	1.2	58.8	0.0	0.0

E-mail: vincze@neb.com



PacBio motif stats

Job ID: 21717
Specificity: ACCNNNNNCTGA
IPD threshold: 2
Modifications source: GFF (fast) CSV (slow)
Calculate

Finding ACCNNNNNCTGA sites in ER2796
762 sites found
Loading modifications.gff
Processing ER2796
Total sites: 762

% of sites with IPD >= 2:	75.5	0.3	0.0	0.3	0.0	0.1	0.1	0.3	0.3	0.1	0.0	0.0
average IPD:	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	A	**C**	**C**	**N**	**N**	**N**	**N**	**N**	**C**	**T**	**G**	**A**
	T	**G**	**G**	**N**	**N**	**N**	**N**	**N**	**G**	**A**	**C**	**T**
average IPD:	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.7	0.0	0.0
% of sites with IPD >= 2:	0.1	0.3	0.0	0.7	0.1	0.1	0.1	0.0	0.0	77.7	0.1	0.0

E-mail: vincze@neb.com

PacBio motif stats

Job ID: 21717
Specificity: ACCNNNNNATGA
IPD threshold: 2
Modifications source: GFF (fast) CSV (slow)
Calculate

Finding ACCNNNNNATGA sites in ER2796
465 sites found
Loading modifications.gff
Processing ER2796
Total sites: 465

% of sites with IPD >= 2:	100.0	1.9	0.2	0.6	0.6	0.0	0.2	0.0	0.0	0.0	0.0	0.0
average IPD:	5.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	A	**C**	**C**	**N**	**N**	**N**	**N**	**N**	**A**	**T**	**G**	**A**
	T	**G**	**G**	**N**	**N**	**N**	**N**	**N**	**T**	**A**	**C**	**T**
average IPD:	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.7	0.0	0.0
% of sites with IPD >= 2:	0.2	0.2	0.4	1.3	0.2	0.0	0.0	0.0	0.2	99.4	0.2	0.0

E-mail: vincze@neb.com

Figure 69: NEB software analysis of SMRT sequencing (*Pacific Biosciences*) data. from the MS fusion modified *E. coli* genomic DNA.

These results indicated that although the enzyme seemed to accept cytosine or thymine at the 9th position, it strongly preferred adenine or guanine (R). This confirms that the DNA recognition sequence for the MS fusion was the same as the wild-type enzyme (ACCN$_5$RTGA).

Reports for Job Ed_1_Dryden_MODs

PACIFIC BIOSCIENCES

SMRT Cells: 1 Movies: 1

Motif Summary

Motifs	Modified Position	Type	% Motifs Detected	# Of Motifs Detected	# Of Motifs In Genome	Mean Modification QV	Mean Motif Coverage	Partner Motif
ACCNNNNNRTGA	1	m6A	99.69%	971	974	89.04	57.17	TCAYNNNNNGGT
TCAYNNNNNGGT	3	m6A	99.69%	971	974	90.00	57.86	ACCNNNNNRTGA
ACCNNNNHRTGGB	1	m6A	49.07%	291	593	54.17	60.71	VCCAYDNNNNGGT
VCCAYDNNNNGGT	4	m6A	45.36%	269	593	54.62	61.85	ACCNNNNHRTGGB
ACCNNNNHRAGA	1	m6A	41.75%	200	479	48.38	61.76	
HTCARNNNNNGGTNV	4	m6A	36.31%	264	727	51.22	62.33	
ACCNNNNNYTGAD	1	m6A	34.9%	320	917	49.93	60.88	

Figure 70: Results from SMRT sequencing (*Pacific Biosciences*) of CC398-1 modified *E. coli* genomic DNA. Table to show detection frequency of modified motifs.

Data collected from the SMRT sequencing of the wild-type (CC398-1) modified DNA (Fig. 70), supports the legitimacy of that collected from the genomic DNA modified by the MS fusion. Multiple DNA motif results in both cases suggests that the results from the MS fusion are reliable. There is also a similar incidence of error in the results from the wild-type. Once again, the complementary sequence to the 5th motif in the list would not contain the requisite adenine base at the 5' end. Again, as this error happens in both cases, it suggests a limitation in the experiment, and not unusual activity in the MS fusion enzyme.

Since it was determined by the *in vivo* assays and SMRT sequencing that the MS Fusion produced an active enzyme, it was considered worthwhile to make further attempts to express and purify the protein. The MS fusion was an unnatural construct and therefore potentially more fragile than the wild-type enzyme. Therefore, a slower expression of the gene could help the protein product remain soluble. An overnight expression (~18 hours) at 20 ^{o}C was conducted on a small scale and the cell-free extract from these cells was then analysed for signs of soluble MS fusion protein (Fig. 71).

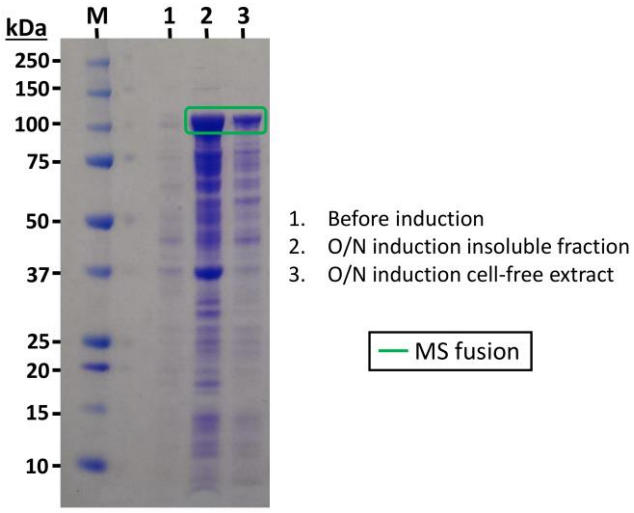


Figure 71: SDS-PAGE gel of a small scale induction of the MS fusion gene and subsequent solubility determination of the MS fusion protein.

Analysis of the cell-free extract showed signs of soluble MS fusion, indicating that it could be expressed on a larger scale and purified. A bigger cell culture (4 L of LB media) was produced, and the gene was induced under these new conditions. The cells were lysed and centrifuged, and the cell-free extract was passed through a nickel affinity column. SDS-PAGE analysis of samples from this purification identified that the MS fusion protein was successfully adhering

to the nickel and that a large proportion of it could be eluted from the column with 500 mM imidazole (Fig. 72).

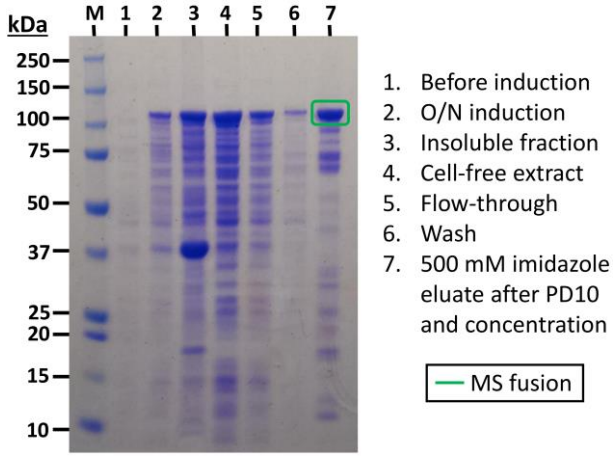


Figure 72: SDS-PAGE analysis of the nickel affinity purification of the overnight expression of the MS fusion protein.

The concentration of the partially purified sample was estimated by UV/vis spectroscopy and then used in a plasmid cleavage assay to identify restriction activity. Work with the EcoKI MS fusion indicated that this CC398-1 MS fusion protein would require the addition of stoichiometric amounts of its M subunit in order to be active. In this preliminary assay, no extra M subunit was added.

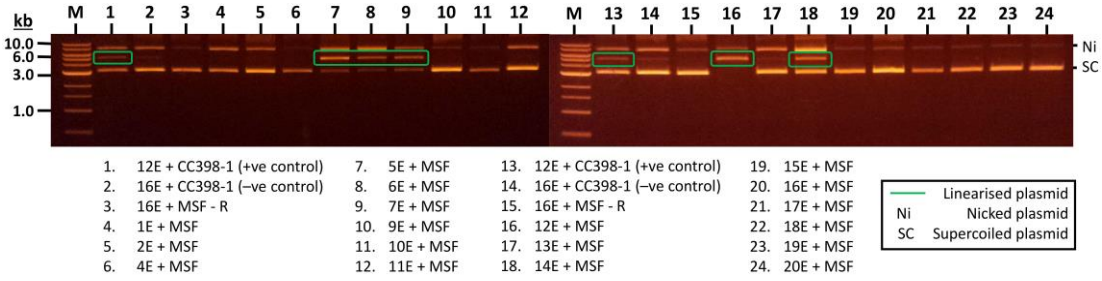


Figure 73: Plasmid cleavage assay of the MS fusion protein.

Results from the Eddy cleavage assay were surprising (Fig. 73). The CC398-1 MS fusion protein appeared to be restriction active, without any supplementary protein. Data collected from SMRT sequencing suggested that this enzyme would show the same pattern of cleavage as the wild-type enzyme (5E, 6E, 7E, 12E and 14E positive), and this indeed was the case. A negative result in the control sample (MS fusion sample without the addition of the R' subunit) (Lane 3), suggested that the positive result was not due to background activity from contaminants. The observed activity was due the MS fusion protein. However, the data from

EcoKI MS fusion suggested that this would not be possible and as such, it was believed that the complementary HsdM had to have been co-purified and retained in the MS fusion sample. This led to the conclusion that the CC398-1 MS fusion protein should undergo further purification to separate it from any potential proteolysis product, such as free HsdM protein.

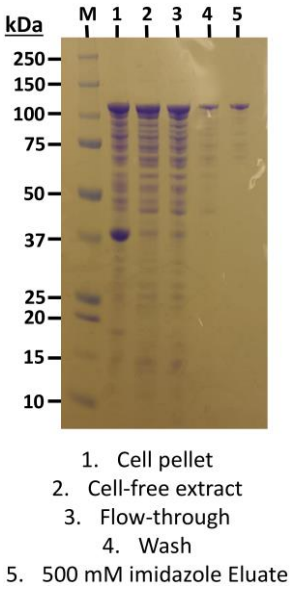


Figure 74: SDS-PAGE analysis of the nickel affinity purification of the MS fusion protein.

A fresh, large-scale culture of cells containing the MS fusion gene was created and lysed. The cell-free extract was purified by nickel affinity (Fig. 74) as before, but then subjected to size exclusion chromatography. The resolution of the S200 gel filtration column meant that any HsdM protein (~59 kDa) co-eluting from the nickel column, would be separated from the much larger MS fusion (~106 kDa).

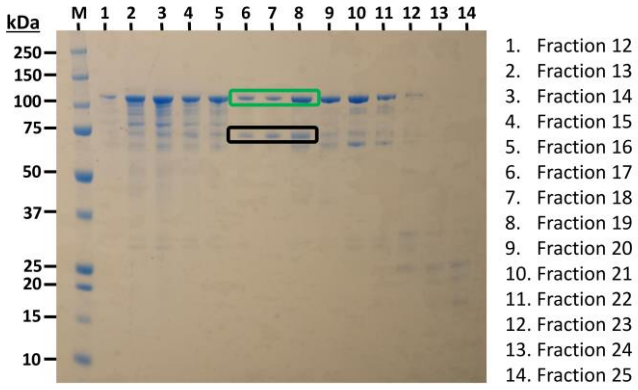


Figure 75: SDS-PAGE analysis of the size exclusion purification of the MS fusion protein. The MS fusion protein is highlighted by the green box and the contaminant is highlighted by the black box.

SDS-PAGE analysis of samples from the size exclusion purification of the MS fusion protein sample showed that there was still a substantial amount of the contaminant (Fig. 75, black box). This suggested that it has a strong affinity for the MS fusion protein, and as such would most likely be some form of the HsdM subunit. With this in mind, the sample was used in a plasmid cleavage assay, again without addition of supplementary HsdM protein.

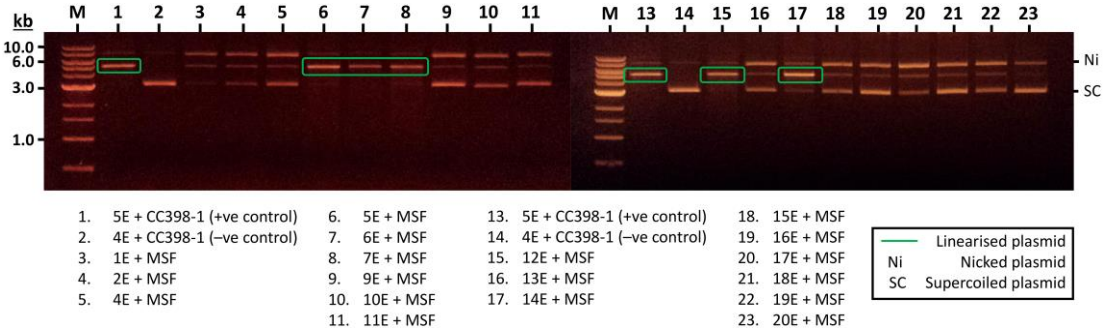


Figure 76: Gel electrophoresis analysis of samples from a plasmid cleavage assay of the MS fusion protein, purified by size exclusion.

Results from the Eddy cleavage assay showed that the MS fusion was indeed active in the same way as before (Fig. 76). The MS fusion was also used in a cleavage assay against modified and unmodified *E. coli* genomic DNA (Fig. 77). Results showed that the fusion protein was restriction active against unmodified (U/M) genomic DNA (Lanes 5 and 6) and λ phage DNA (Lanes 9 and 10), and was inactive against self-methylated genomic DNA (Lanes 7 and 8).

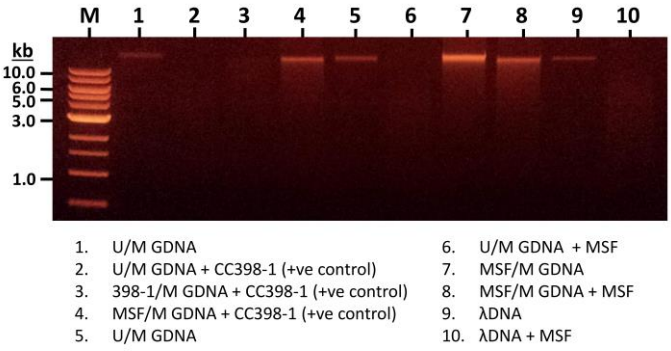


Figure 77: Gel electrophoresis analysis of samples from a genomic DNA (GDNA) cleavage assay of the MS fusion protein. Lane 1 contains unmodified (U/M) genomic DNA alone. Lane 2 shows that with the addition of wild-type CC398-1 MTase (+ve control), the genomic DNA no longer appears on the gel. 398-1/M GDNA denotes genomic DNA that has been modified by the wild-type enzyme. MSF/M denotes genomic DNA that has been modified by the MS fusion. Lanes 9 and 10 contain lambda **(λ)** phage DNA.

The purified MS fusion protein was also subjected to analysis by gel filtration HPLC. If the MS fusion were purified without the extra fragment, it is reasonable to assume that it would be unable to take on the quaternary structure of the wild-type enzyme. As the M and S subunits are fused, if the protein were to dimerise, it would form an M_2S_2 structure. However, evidence suggests that the contaminant is some form of the HsdM, and is allowing the enzyme to take on a pseudo M_2S_1 conformation. Gel filtration data would give an indication as to which structure is most likely.

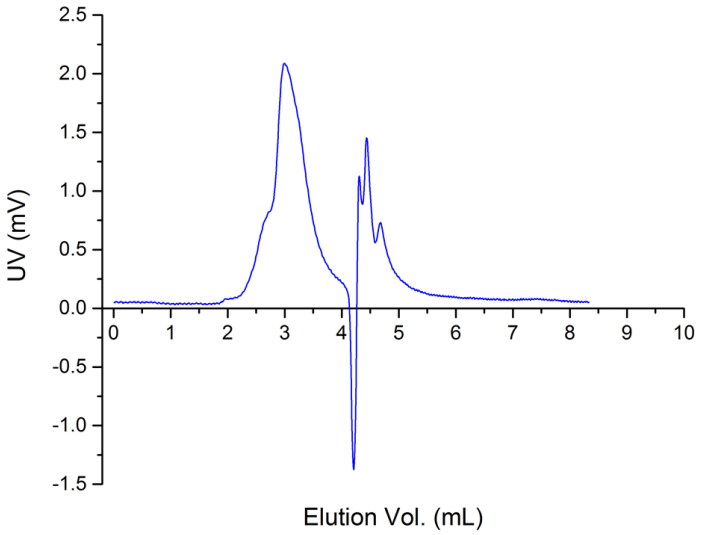


Figure 78: Gel filtration HPLC elution profile of the MS fusion protein.

Data received from the gel filtration HPLC analysis of the MS fusion was reasonably clear (Fig. 78). Due to the unstable nature of the enzyme, it was only possible to purify the protein partially. As such, it was not expected that the protein solution would produce a neat elution profile. There was a shoulder to left side of the main peak occurring at an elution volume of 2.99 mL, suggesting a larger contaminant. There were three smaller peaks at a higher elution volume, but were due to glycerol and small contaminants. As the elution volume of the CC398-1 wild-type was 3.1 mL, a smaller elution volume for the MS fusion indicates that it is indeed larger. The poor resolution of these complete MTase complexes in gel filtration means that it was not possible to get an accurate mass from these data. However, the values can be compared relatively. Using the obtained elution volumes, the calculated mass of the MS fusion is 311.5 kDa, whilst the calculated mass of the wild-type was 210 kDa (actual mass, 166 kDa). The extra fragment had a molecular weight of approximately 60 kDa (estimated from SDS-PAGE analysis), and so was still the most likely cause of this increase in mass and quicker elution time.

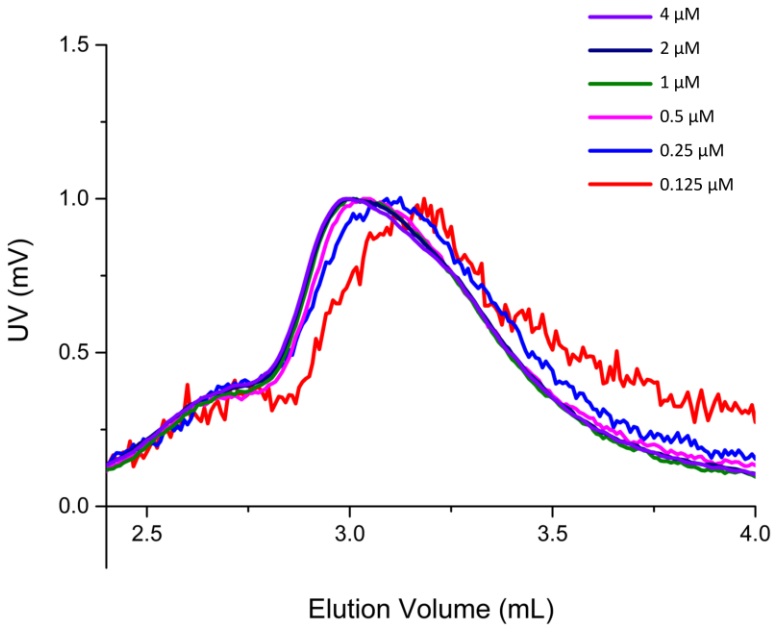


Figure 79: Normalised elution profiles of decreasing concentrations of the MS fusion.

As with the wild-type MTase, HPLC elution profiles of decreasing concentrations of the MS fusion protein gave an indication of the stability of this enzyme complex (Fig. 79). The pseudo M_2S_1 form of the protein appeared to start breaking down to any significant extent at a concentration of 0.25 µM. Although it should be appreciated that this concentration is not definite (due to the level of impurity of the protein solution), it did seem that the MS fusion showed the same level of stability as the wild-type enzyme. Additionally, if the concentration were inaccurate, it would be lower than calculated. This would mean that this MTase was stable at lower concentrations than the wild-type.

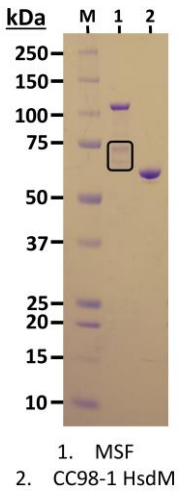


Figure 80: SDS PAGE gel of purified MS fusion and HsdM samples. The contaminating species is highlighted by the black box.

Comparison of SDS-PAGE gel bands of the purified MS fusion and HsdM subunit samples indicated that the contaminant was not exactly the same species as the HsdM (Fig. 80). The gel showed two different bands (black box), both of a higher molecular weight than the HsdM protein. If the fragment is an MS fusion proteolysis product, it would also include part of the HsdS sequence. Evidence suggested that the contaminating fragment was binding the fusion protein in a manner that was not inhibiting activity. It can be inferred from this that the fragment was able to fold correctly, suggesting that the peptide sequence had degraded to a specific point in the secondary structure that had maintained its tertiary structure. It was therefore thought most likely that the fragment would contain at least the first TRD of the HsdS. To examine the association between the MS fusion protein and the unknown fragment, a cross-linking experiment was carried out. The fusion protein sample was incubated with glutaraldehyde and examined by SDS-PAGE. If the unknown fragment was closely associated with the MS fusion protein, the incubation would result in a cross-linked product of the two species, and would therefore appear at a higher molecular weight on the SDS-PAGE gel.

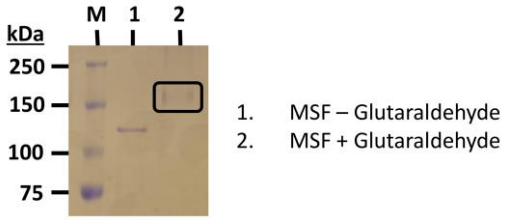


Figure 81: MS fusion sample ± treatment with glutaraldehyde.

Comparing plus and minus glutaraldehyde samples identified that the MS fusion species in the glutaraldehyde positive sample had increased in molecular weight (Fig. 81). It could also be observed that the increase could be attributed to the molecular weights of either of the unknown fragments (60 to 70 kDa). This proved that there was a strong association between the MS fusion and the bound fragment, and provided further evidence that the fragment contained the HsdM subunit.

Peptide fragmentation mass spectrometry analysis of this protein was carried out, in order to identify the unknown proteins in the MS fusion sample. An SDS-PAGE gel of the partially purified MS fusion sample was run, and the MS fusion and contaminating bands were excised and sent for analysis (Fig. 82).

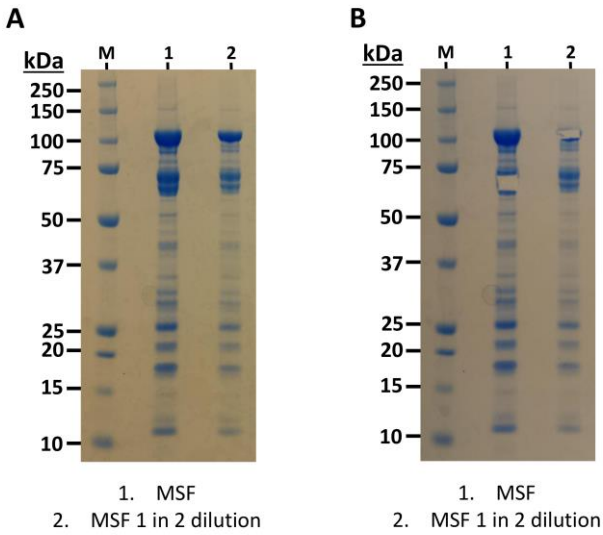


Figure 82: SDS-PAGE gel showing nickel affinity purified MS fusion (Lane 1) and a 2X dilution (Lane 2) (A). The gel after band excision (B).

The data retrieved from analysis by mass spectrometry identified three separate species. The first, from the band of highest molecular weight was confirmed as the MS fusion protein. The ~70 kDa molecular weight protein was identified as a contaminating protein from *E. coli* (ArnA), whilst the smallest protein species (~60 kDa) contained peptides from both HsdM and HsdS subunits, as expected (See Appendix F for Mass spectrometry results). It was concluded from this data that the protein fragment associating with the MS fusion was a product of proteolysis.

Evidence suggested that the bound fragment was not only relatively stable and strongly associating with the MS fusion protein, but was also inducing the MS fusion activity. The stability of this fragment led to the belief that it would include the first TRD of the HsdS subunit, in order for it to fold correctly. The estimated molecular weight of the MS fusion, up to the C-terminal of the first TRD is ~81 kDa. However, by comparing SDS-PAGE gels of the MS fusion protein, the fragment does not appear large enough (<65 kDa). Although there was not total peptide sequence coverage of the protein fragment, there were hits across the full length of the HsdM. This was not the case for the HsdS, and as such there was relatively low confidence in the protein match (a score of 120, compared to 1271 for the HsdM). If the final peptide hit on the HsdS occurs at the end or near the end of the fragment, the molecular weight of this fragment would be ~ 73 kDa. This too seems too large when compared to the molecular weight estimated from the SDS-PAGE gel. If the fragment has a molecular weight of approximately 65 kDa, then the first peptide hit on the HsdS is a good candidate for the end

of the fragment. The estimated molecular weight of the MS fusion protein, truncated to the end of this peptide match is 62 kDa.

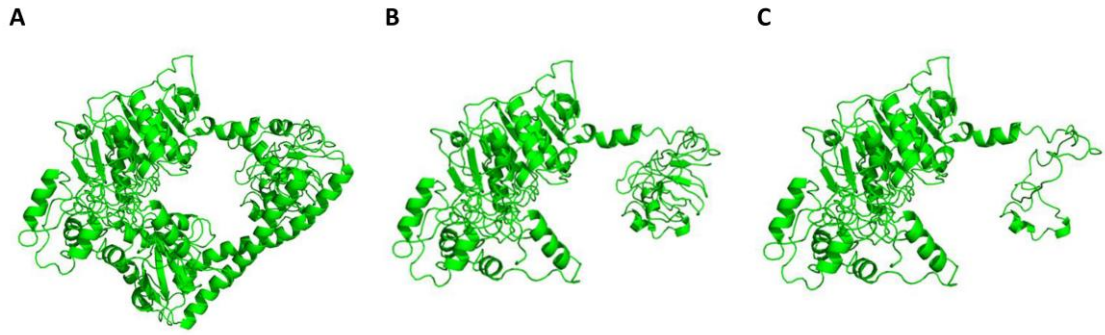


Figure 83*:* Cartoon model of the MS fusion protein (A). Cartoon model of the contaminating fragment, truncated to the final peptide mass spectrometry hit on the HsdS (B). Cartoon model of the contaminating fragment to the first peptide hit on the HsdS (C)

Matching the experimental data from the fragment to the model of the MS fusion raised further questions. Using both estimates for the molecular weight of the fragment created two distinct models. Enzyme assays showed that the fragment was stabilising the MS fusion and enabling its activity, yet neither estimate included the complete first TRD (Fig. 83B and C). The structure in the smaller estimate (Fig. 83C) contains only four residues after the N-terminal region. As this is the section of the HsdS that is proposed to associate with the HsdM subunit, this is the estimate that seems most likely. However, the MS fusion contains the complete HsdS and as such would be associating with the HsdM in the extra fragment. This could possibly prevent the N-terminal sequence on the fragment from binding. If this is the case, the extra sequence might be moving freely. This again makes the shorter sequence more likely. Nevertheless, it was clear that this C-terminal sequence was not affecting its association with the MS fusion protein.

3.6. CC398-1 HsdM Half HsdS

The HsdS of a Type I MTase contains two TRDs, which each correspond to one half of the bipartite DNA recognition sequence. These TRDs are joined by an amino acid sequence, which is highly conserved within its enzyme family, and is therefore known as the central conserved region. It is proposed that the Type I MTase has a symmetrical organisation, where the C-terminal of the HsdS loops around to meet its N-terminal (Kneale 1994). Work from several research groups provided evidence for this circular model by creating mutants of the HsdS, where each mutant possessed a different C-terminal but retained wild-type specificity (Taylor et al. 1994; Janscak & Bickle 1998). Investigations using the Type IC enzyme, EcoDXXI, also found that removing either the N-terminal or C-terminal TRD gives rise to a new DNA specificity (MacWilliams & Bickle 1996). Given this evidence, it was thought that the HsdS of the CC398-1 MTase could be truncated to the end of the first TRD, and that this would cause the structure to form a homodimer and therefore recognise a palindromic DNA sequence (Fig. 84B).

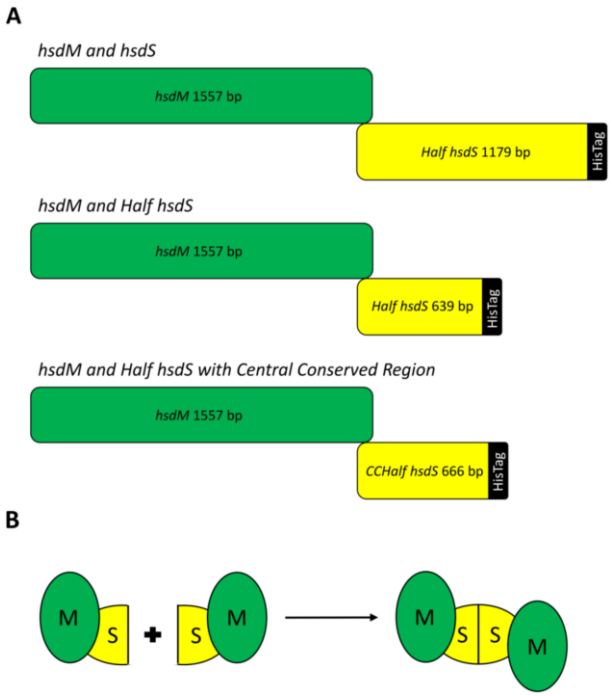


Figure 84: Cartoon diagram of the genetic arrangement of the two Half HsdS constructs (A), and the proposed protein subunit conformation (B).

Two different half HsdS constructs were designed from the CC398-1 MTase template (Fig. 84A). One of these ended directly after the estimated end of the helical spacer, whilst the other incorporated the rest of central conserved region. It was thought that only by including the entire sequence up to the second TRD, would the enzyme subunit be able to dimerise and form

a circular structure. The first structure was named “HalfSHis”, whilst the second was known as “CCHalfSHis” or “CHS”, as it contained all of the central conserved region (Fig. 85). These two structures were created by PCR, and subsequently ligated into the pJF vector (see Appendix C for plasmid maps). Under normal circumstances, proteins produced from this vector possess a C-terminal HisTag, hence the suffix “His”.

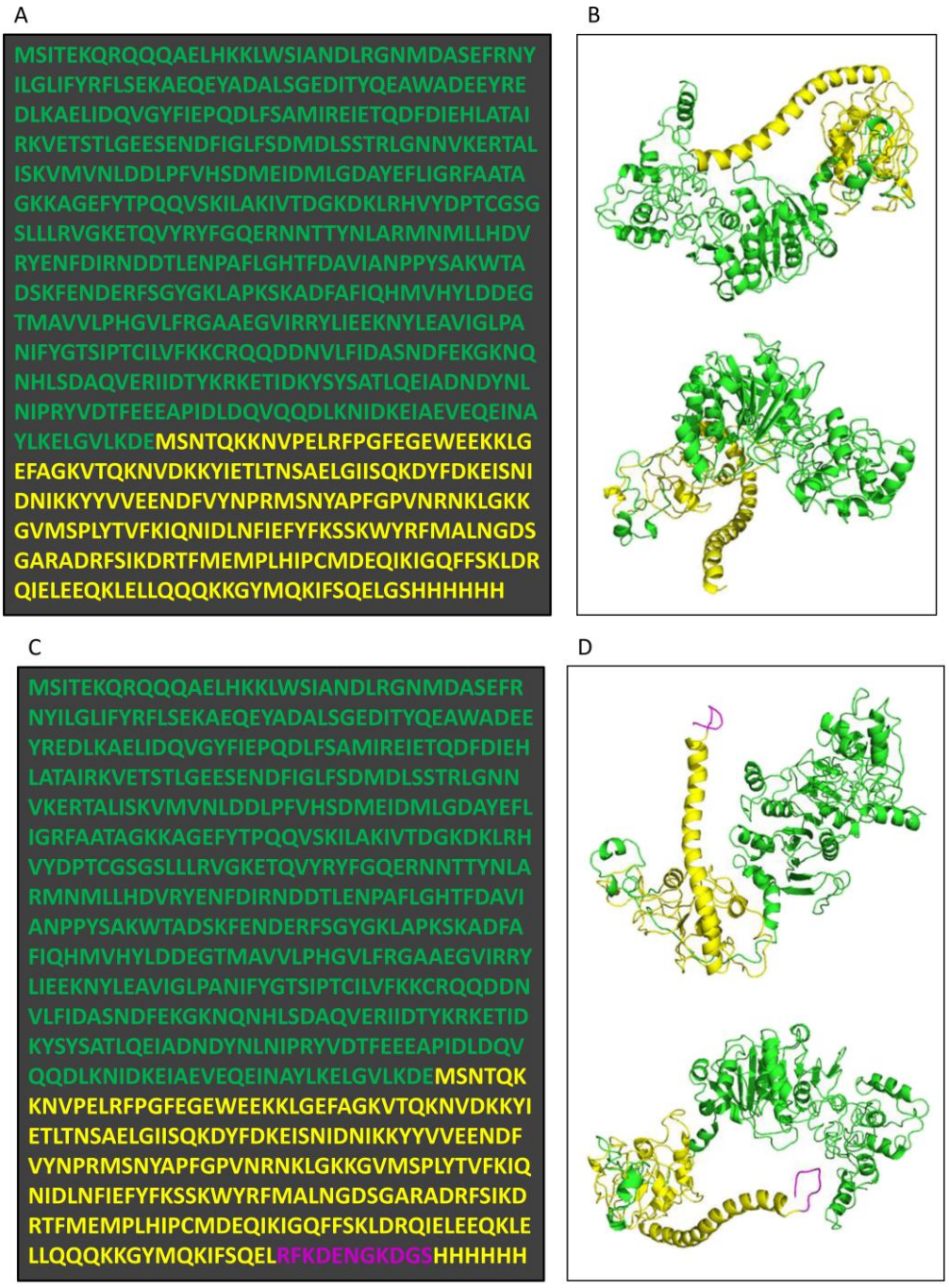


Figure 85: Amino acid sequence of the HalfSHis MTase (A) Model of the HalfSHis MTase (B). Amino acid sequence of the CCHalfSHis MTase (C) Model of the CCHalfSHis MTase (D). The extra residues in the CCHalfSHis sequence are coloured purple.

The S subunit of the HalfSHis enzyme had a predicted molecular weight of 25896.5 Da. The CCHalfSHis construct had an extra nine residues (Fig. 85. highlighted in purple) and a predicted molecular weight of 26986.6 Da.

Small scale tests showed that both of the halved HsdS constructs could be expressed and were soluble. Expression was then carried out on a larger scale, in order to purify an amount of the proteins that could be analysed. First produced was the "HalfSHis" enzyme, which was subsequently purified by nickel affinity chromatography (Fig. 86).

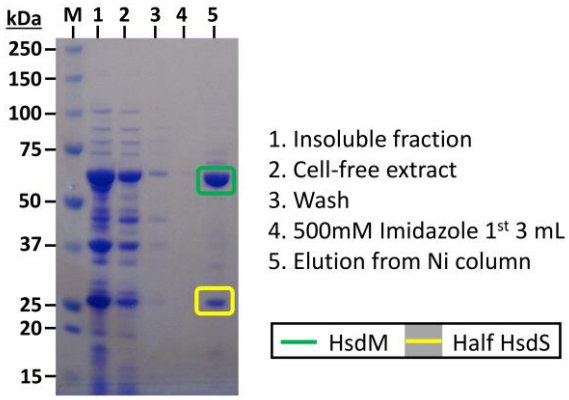


Figure 86: SDS-PAGE analysis of samples from the nickel affinity purification of HalfSHis.

The HalfSHis enzyme appeared to purify well after one chromatography step, and so was then assayed for restriction activity with a plasmid cleavage assay (Fig. 87). As the activity of this novel construct was unknown at this stage, it was first checked against positives for the wild-type enzyme, Eddys 5E and 7E.

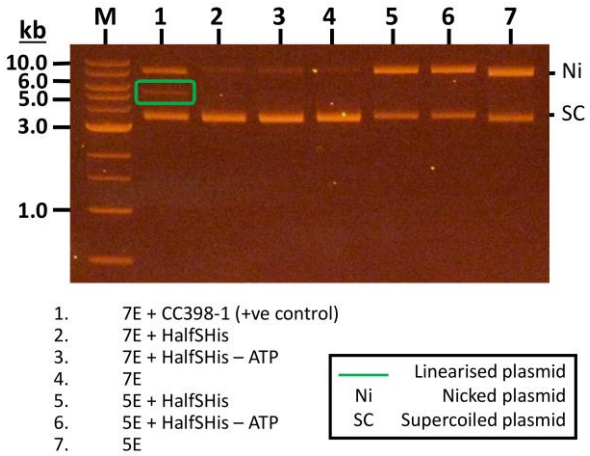


Figure 87: Plasmid cleavage assay of the HalfSHis protein.

No signs of restriction activity were found in the preliminary test. At this juncture, the other half HsdS construct, CCHalfSHis, was purified by nickel affinity. A further size exclusion chromatography step was added, to reduce the risk of inhibition by contaminants. The gene expressed and seemed to produce a large amount of the target protein. The recombinant protein was successfully purified by nickel affinity, and further by size exclusion (Fig. 88).

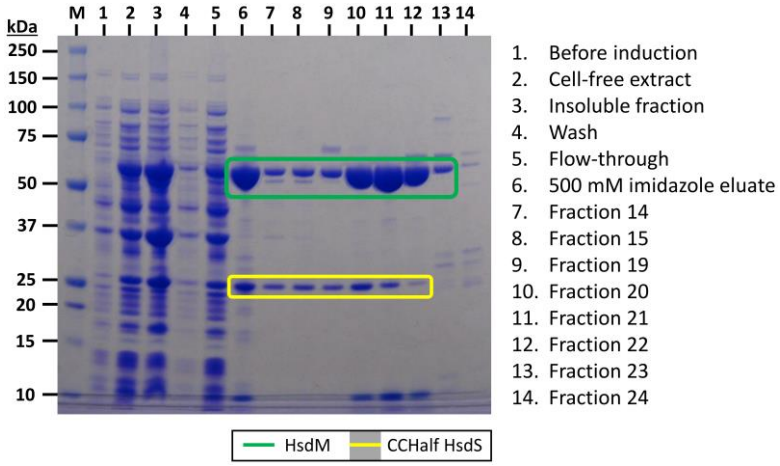


Figure 88: SDS-PAGE analysis of samples from the purification of the CCHalfSHis protein.

The lack of restriction activity in HalfSHis was thought to be due to the fact that the specificity of the enzyme was not yet known. Therefore, a wider range of plasmid species were tested in a new cleavage assay of CCHalfSHis.

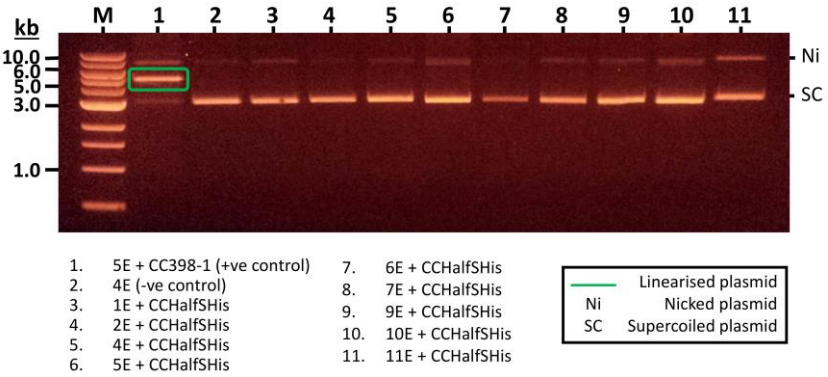


Figure 89: Plasmid cleavage assay of the CCHalfSHis protein.

As with the HalfSHis assay, CCHalfSHis produced no linearised species in the plasmid cleavage assay (Fig. 89). It was assumed that to become active, the Half HsdS proteins would have to dimerise. With this in mind, it was proposed that the HisTag on the end of this subunit could be interfering with this process. It was therefore decided that the truncated MTase should be expressed and purified without the tag.

At this stage, CCHalfSHis was considered the most likely to produce an active enzyme, due to its inclusion of the entire central conserved region. It was therefore the species used for investigating the effects of removing the HisTag. A stop codon was introduced to the end of the *cchalfS* gene by PCR, and the product was ligated into the pJF vector. The gene was then expressed and the solubility of the protein product was verified by SDS-PAGE. The gene was then expressed in a larger amount of medium, in order to purify the protein. As the pI of the CCHalfS protein was estimated to be 5.05, the cell-free extract of the subsequent cell culture was first passed through an anion exchange column, and then a gel filtration column. A final purification step was carried out, using a heparin agarose column. Heparin agarose binds to DNA-binding proteins, which can then be eluted with an increasing salt gradient. Samples from all of these purification steps were analysed by SDS-PAGE (Fig. 90).

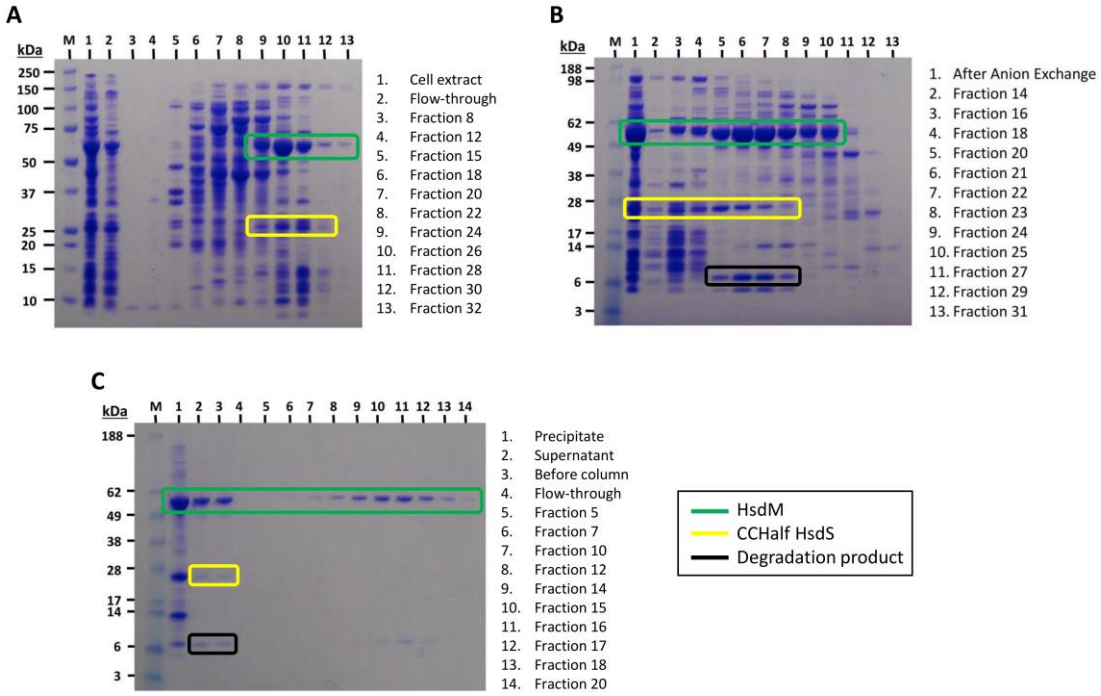


Figure 90: SDS-PAGE analysis of samples from the purification of the CCHalfS protein, without HisTag. In chronological order, anion exchange (A), gel filtration (B) and heparin agarose (C).

The SDS-PAGE gels showed that the purification of the CCHalfS protein was unsuccessful. Successive purification steps saw a reducing amount of the Half HsdS subunit and an increase in a smaller band on the gels (Fig. 90, highlighted by the black box), which appeared to be a product of the degradation of the half HsdS. The subunit was clearly unstable under these conditions.

As with the MS fusion protein, it was considered more effective to test the activity of the half hsdS enzymes in an *in vivo* assay. This would confirm whether the constructs were viable and postpone the need to produce large amounts of soluble, stable protein.

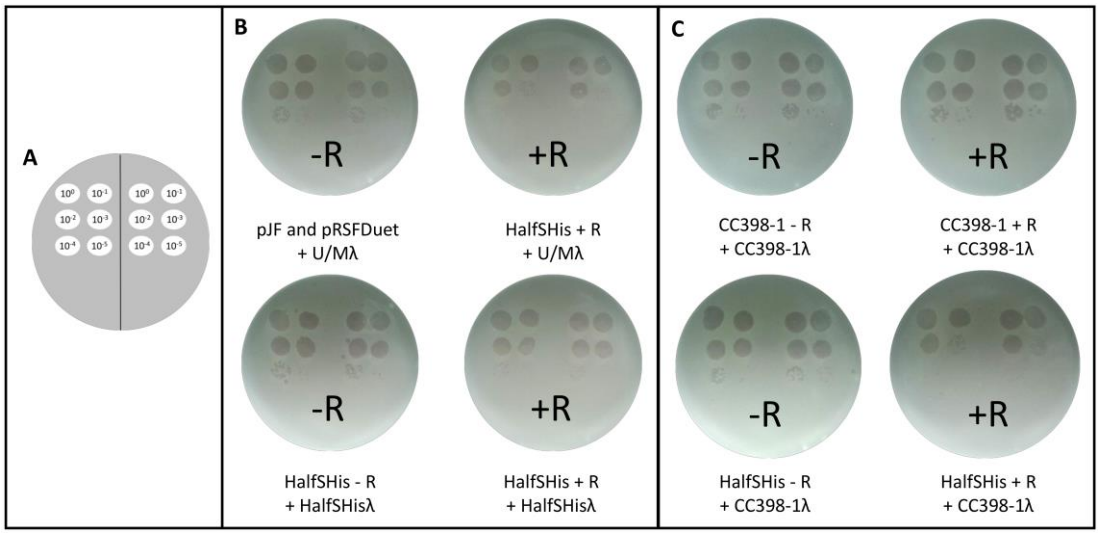


Figure 91: *In vivo* spot tests to assay for R-M activity in the HalfSHis protein. Diagram of spot test dilutions (A) The HalfSHis MTase in restriction complex, against unmodified (U/Mλ) and HalfSHis Modified λ phage (HalfSHisλ) DNA (B) Wild-type CC398-1 control and the HalfSHis in restriction complex. against CC398-1 modified (CC398λ) λ phage DNA (C).

The spot test *in vivo* assays of the HalfSHis MTase were successful (Fig. 91). When tested against unmodified λ phage DNA, the HalfSHis (with supplementary R subunit) showed clear restriction activity, to at least one log dilution. The MTase also appeared to be modification active, as the restriction complex seemed unable to restrict the phage that had been passed through the same system (Fig. 91B). Yet more encouraging was that the HalfSHis MTase was able to restrict phage that had been modified by the wild-type enzyme (Fig. 91C). This was good evidence that the new construct had a new DNA recognition sequence. To verify these preliminary results, full plate *in vivo* assays of the HalfS MTase were carried out.

HalfSHis MTase *In vivo* Assay Results:

Phage Type	Phage Dilution	Phage Volume	R-M System	Number of Plaques
Unmodifiedλ	10^{-5}	22.5 µL	None	384
	10^{-4}	27.8 µL	HalfSHis + R	1097
	10^{-6}	100 µL	None	192
	10^{-5}	286 µL	HalfSHis + R	1270
	10^{-6}	100 µL	None	274
	10^{-5}	40 µL	HalfSHis + R	512
	10^{-5}	100 µL	None	1516
	10^{-6}	40 µL	HalfSHis + R	36
	10^{-6}	100 µL	None	263
	10^{-5}	40 µL	HalfSHis + R	493
HalfSλ	10^{-5}	28.5 µL	HalfSHis	264
	10^{-5}	35.7 µL	HalfSHis + R	273
	10^{-6}	167 µL	HalfSHis	185
	10^{-6}	334 µL	HalfSHis + R	253
	10^{-6}	200 µL	HalfSHis	311
	10^{-6}	200 µL	HalfSHis + R	215
	10^{-6}	200 µL	HalfSHis	324
	10^{-6}	200 µL	HalfSHis + R	229
	10^{-5}	20 µL	HalfSHis	343
	10^{-5}	20 µL	HalfSHis + R	221

Table 4: Raw data from full plate *in vivo* assays of the HalfSHis MTase. The table shows ±R pairs, which are the results from the experiment (+R) and control (-R), and the subsequent repeats. The plaque numbers cannot be compared without adjusting for volume and dilution.

The E.O.P. of the unmodified λ phage against the restriction active HalfSHis MTase was 0.40 ±0.20 (see Appendix E for calculations). Comparing this value to those gathered from the wild-type and MS fusion enzymes (0.15 and 0.09 respectively), it suggests that the HalfS is less active in a restriction complex. However, due to the large margin of error in this assay, this value should be considered the same as wild-type. The E.O.P. of the HalfSHis modified λ phage was 0.70 ±0.20. This indicates that the HalfSHis has similar MTase activity to both the MS fusion (0.78) and the wild-type (0.88) enzymes. This was quite surprising given that this construct does not possess the whole sequence of the central conserved region.

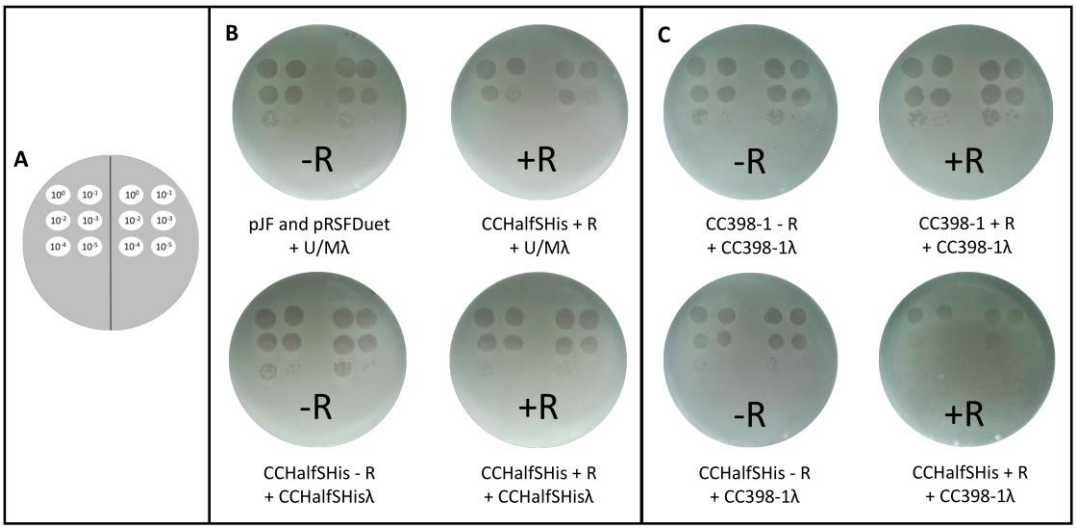


Figure 92: *In vivo* spot tests to assay for R-M activity in the CCHalfSHis protein. Diagram of spot test dilutions (A) The CCHalfSHis MTase in restriction complex, against unmodified (U/Mλ) and CCHalfSHis modified (CCHalfSHisλ) λ Phage DNA (B) Wild-type CC398-1 control and the CCHalfSHis in restriction complex. against CC398-1 modified (CC398-1λ) λ phage DNA (C).

The same procedure was carried out using the *cchalfS* gene in the pJF vector, and this too was a success. When tested against unmodified λ phage DNA, the CCHalfSHis (with supplementary R subunit) was restricting the phage by at least one log dilution. Modification activity was identified, as this MTase in restriction complex was also unable to restrict the phage that had been passed through the same system (Fig. 92B). As with the HalfSHis MTase, the CCHalfSHis was able to restrict CC398-1 modified phage, and therefore also had a different DNA recognition sequence from wild-type. These results were interesting. From the *in vivo* work, it appeared that the two half HsdS constructs had similar activity, despite one of them lacking the central conserved region. This result was investigated further by using full plate *in vivo* assays of the CCHalfSHis protein.

CCHalfSHis MTase *In vivo* Assay Results:

Phage Type	Phage Dilution	Phage Volume	R-M System	Number of Plaques
Unmodifiedλ	10^{-5}	22.5 µL	None	384
	10^{-4}	16 µL	CCHalfSHis + R	1164
	10^{-6}	100 µL	None	192
	10^{-5}	84 µL	CCHalfSHis + R	716
	10^{-6}	100 µL	None	274
	10^{-5}	40 µL	CCHalfSHis + R	575
	10^{-5}	100 µL	None	1516
	10^{-6}	40 µL	CCHalfSHis + R	62
	10^{-6}	100 µL	None	263
	10^{-5}	40 µL	CCHalfSHis + R	596
CCHalfSHisλ	10^{-6}	96 µL	CCHalfSHis	294
	10^{-6}	286 µL	CCHalfSHis + R	616
	10^{-6}	100 µL	CCHalfSHis	360
	10^{-6}	100 µL	CCHalfSHis + R	274
	10^{-6}	100 µL	CCHalfSHis	371
	10^{-6}	100 µL	CCHalfSHis + R	274
	10^{-5}	10 µL	CCHalfSHis	434
	10^{-5}	10 µL	CCHalfSHis + R	339

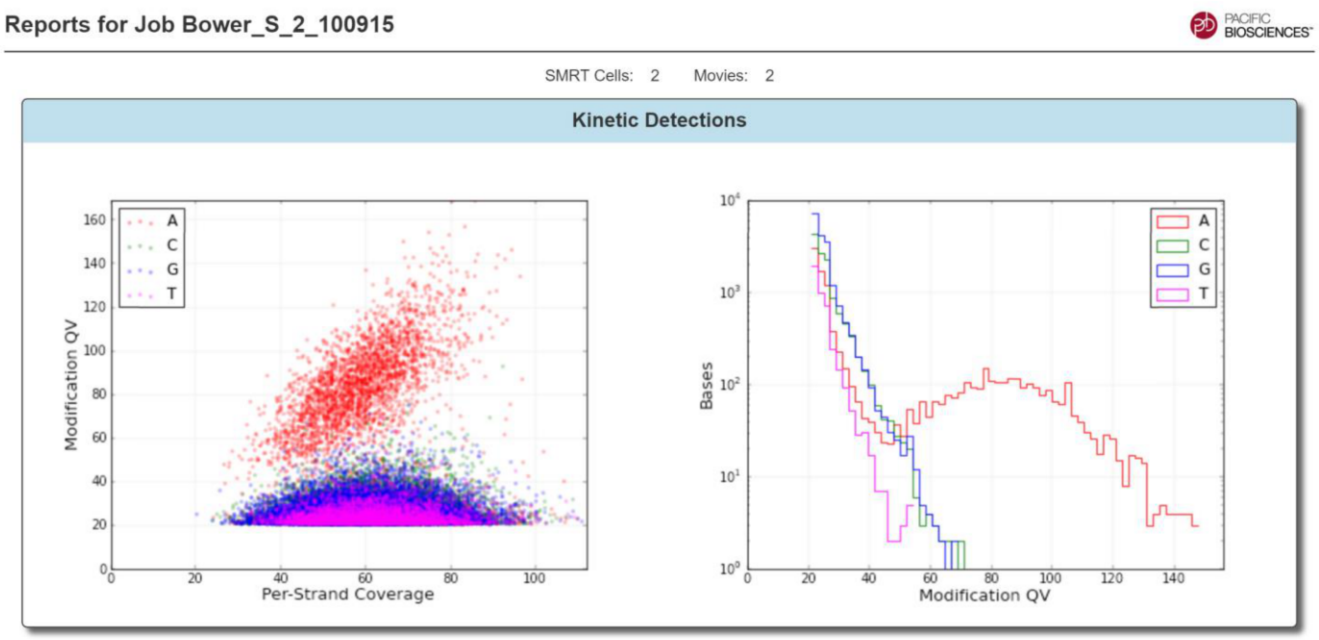
Table 5: Raw data from full plate *in vivo* assays of the CCHalfSHis MTase. The table shows ±R pairs, which are the results from the experiment (+R) and control (-R), and the subsequent repeats. The plaque numbers cannot be compared without adjusting for volume and dilution.

The E.O.P. of the unmodified λ phage against the restriction active CCHalfSHis MTase was 0.58 ±0.21 (see Appendix E for calculations). This is a relatively high value, and suggests that the CCHalfSHis forms a less active restriction complex than any of the previous MTases. Factoring in the error could give a E.O.P of around 0.37. This still seems to lie outside the values obtained with the other MTases but does make it comparable to the HalfSHis MTase. The CCHalfSHis modified λ phage produced an E.O.P. of 0.75 ±0.15. This value lies within the error margins of the other MTases, and suggests this MTase has a modification activity similar to its HalfSHis counterpart.

The two half hsdS genes were able to express enzymes with *in vivo* activity. It was therefore possible to use SMRT sequencing to identify where they were methylating DNA, and therefore determine the new DNA recognition sequences.

HalfSHis and CCHalfSHis MTase SMRT Sequencing Results:

A



B

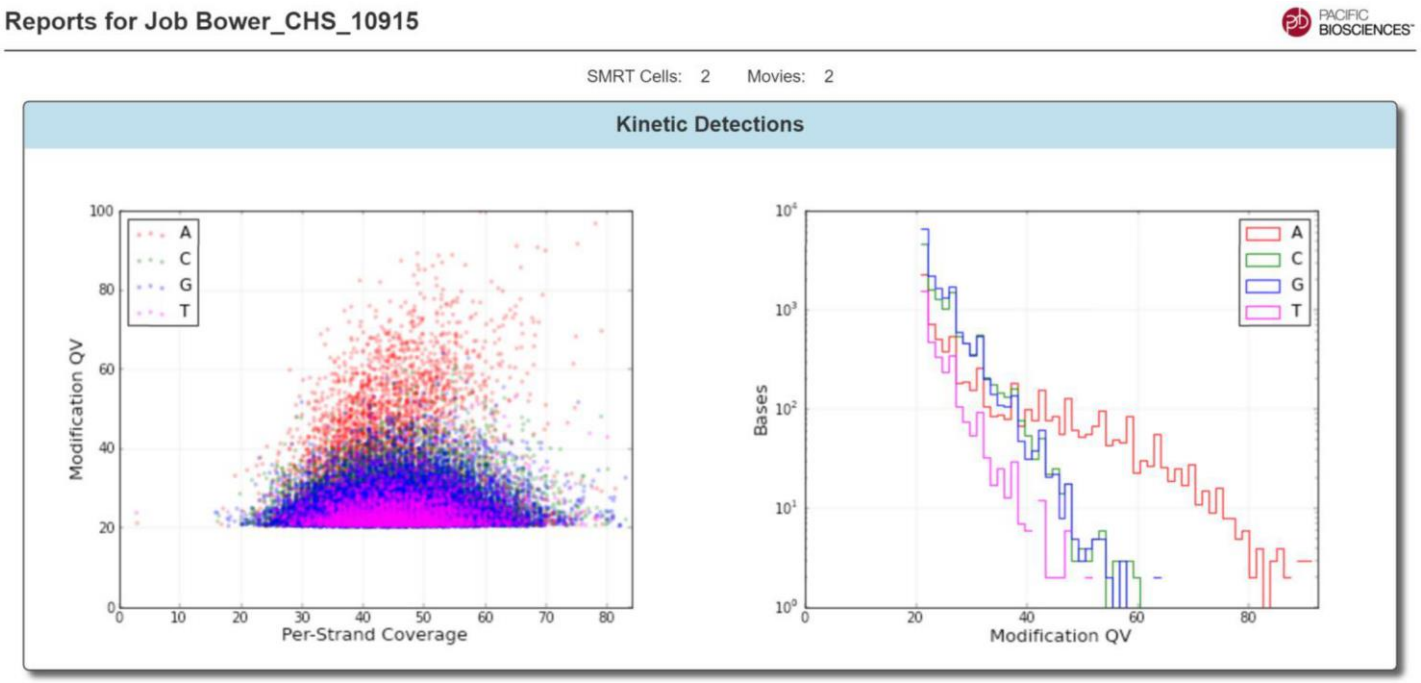
SMRT Cells: 2 Movies: 2

Motif Summary

Motifs	Modified Position	Type	% Motifs Detected	# Of Motifs Detected	# Of Motifs In Genome	Mean Modification QV	Mean Motif Coverage	Partner Motif
ACCNNNNNGGT	1	m6A	99.47%	2650	2664	84.54	58.81	ACCNNNNNGGT

Figure 93: Results from SMRT sequencing (*Pacific Biosciences*) of HalfSHis modified *E. coli* genomic DNA. Kinetic Detections (A). The first graph is a scatter plot showing the detection of base methylation. The higher coverage and detection of adenine methylation (red dots) confirms the HalfSHis protein forms an adenine methylase. The second graph shows detection of bases against detection of methylated bases. This too confirms adenine methylation. Table to show statistical data for the modified motifs (B). The table shows the single motif that was detected, $ACCN_5GGT$.

A



B

SMRT Cells: 2 Movies: 2

Motif Summary

Motifs	Modified Position	Type	% Motifs Detected	# Of Motifs Detected	# Of Motifs In Genome	Mean Modification QV	Mean Motif Coverage	Partner Motif
ACCNNNNNGGT	1	m6A	78.64%	2095	2664	48.22	44.03	ACCNNNNNGGT

Figure 94: Results from SMRT sequencing (*Pacific Biosciences*) of CCHalfSHis modified *E. coli* genomic DNA. Kinetic Detections (A). The first graph is a scatter plot showing the detection of base methylation. The higher coverage and detection of adenine methylation (red dots) confirms the CCHalfSHis protein forms an adenine methylase. The second graph shows detection of bases against detection of methylated bases. This too confirms adenine methylation. Table to show statistical data for the modified motifs (B). The table shows the single motif that was detected, ACCN$_5$GGT.

SMRT sequencing of the HalfSHis modified genomic DNA was successful. After only a single run, the obtained data showed that the HalfSHis MTase was modifying DNA, had a strong preference for modifying adenine bases (Fig. 93A), and modified adenines within a single, distinct DNA sequence motif ($ACCN_5GGT$) (Fig. 93B). The DNA motif is palindromic, which confirms that the new construct is dimerising and supports the idea that the MTase has a circular conformation. If the two HalfSHis units were side by side, the expected motif would be ACCNNNNNCCA. However, as the second half of the bipartite sequence is the complement of this, it suggests the second HalfS unit is recognising the ACC on the anti-sense strand. What was particularly interesting, was that this new palindromic sequence ($ACCN_5GGT$) is a single nucleotide shorter than the wild-type recognition sequence ($ACCN_5RTGA$). What is more, the distance between the methylated bases is one nucleotide further away, relative to the wild-type sequence (See Appendix G for additional results).

The SMRT sequencing of CCHalfSHis modified genomic DNA was also a success, and produced very similar results to the other half HsdS MTase (Fig. 94). This supports the conclusions from the *in vivo* assays, that the two MTase have similar activities. That the CCHalfSHis MTase possessed the same DNA recognition sequence as the HalfSHis species, and that they share the same difference from the wild-type sequence was interesting. The longer primary amino acid sequence of the CCHalfSHis species did not appear to have an effect on recognition sequence length (See Appendix F for additional results).

In vivo assays and SMRT sequencing had proven the half HsdS proteins were active MTases. This suggested that it might be possible to overexpress the genes and purify the subsequent protein, in order to perform work *in vitro*. Previous attempts at this had shown that the enzymes were relatively unstable and that this caused a breakdown of the proteins over time. A faster method of purification ("Quick") was developed, which involved nickel affinity chromatography to separate the MTases from most of the cell extract, and then buffer exchange via a PD-10 desalting column (*GE Healthcare*) (see Materials and Methods), to remove imidazole from the protein solutions. This was successfully performed for both MTases and the results were analysed by SDS-PAGE (Fig. 95).

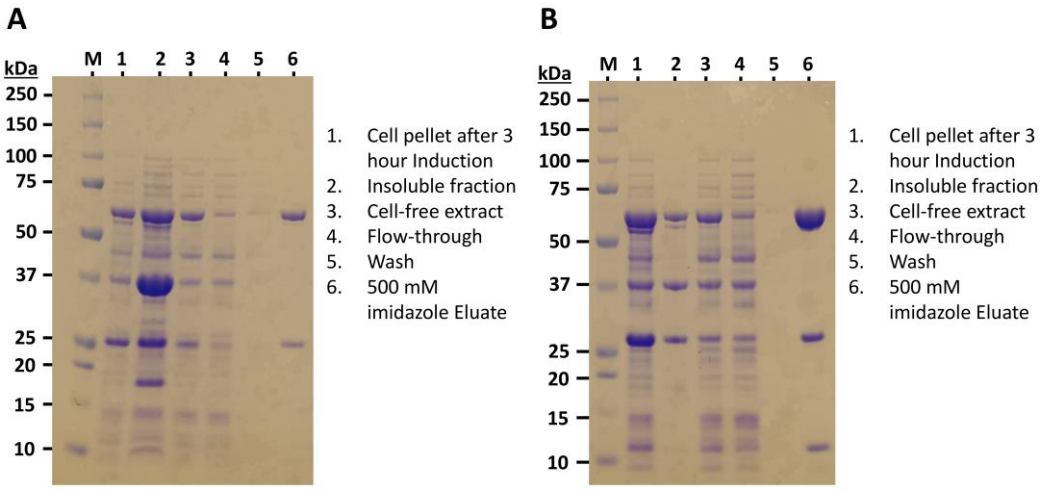


Figure 95: SDS-PAGE analysis of samples from the "Quick" purifications of the HalfSHis (A) and the CCHalfSHis (B) MTases.

The Quick purification yielded more than enough of the two different MTases. 6 g of bacterial cell pellet yielded 25.6 mg of HalfSHis, and 24.1 mg of CCHalfSHis. The vast majority of other proteins had been removed from the protein solutions, leaving them reasonably pure. The calculated yield is therefore fairly reliable.

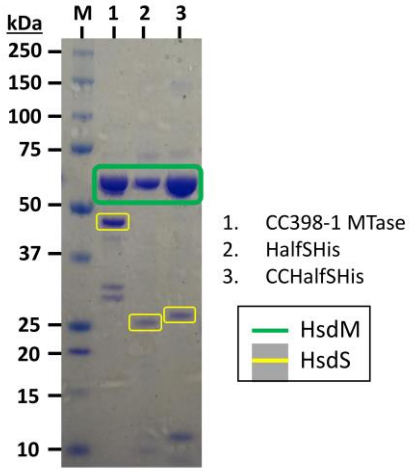


Figure 96: SDS-PAGE gel showing the CC398-1 wild-type MTase, and the two half HsdS proteins.

SDS-PAGE analysis of the two half HsdS enzymes clearly shows their difference in size (Fig. 96). The M subunit remains the same apparent size as the wild-type enzyme, whilst the S subunit has decreased in size. The CCHalfSHis S subunit looks bigger than that of the HalfSHis, corresponding to the predicted difference in molecular weight of 1090.1 Da. Image J analysis of band intensity was again used to provide an impression of the amount of the half

S subunits, relative to the HsdM. The difference in yield between the two MTase subunits was of particular concern, given the degradation of the half S over time. Comparing protein bands in SDS-PAGE gels from four separate purifications of the HalfSHis and CCHalfSHis MTases, provided average band intensities. These were then used to calculate standard deviation in order to gauge the margin of error. The ratio of HsdM to HalfSHis was 1.00 (±0.46):1.04 (±0.72). The ratio of HsdM to CCHalfSHis was 1.83 (±0.79):1.00 (±0.80) (See Appendix D for calculations). In contrast to the wild-type MTase, the two subunits in both of these two half HsdS constructs needed to purify with an even stoichiometry. If the wild-type is M_2S_1, then the half S enzymes should be one M and one $S_{½}$. This supports the Image J results from the HalfSHis MTase, despite the substantial degree of error. If the results from the CCHalfSHis are taken at face value, an explanation for the extra amount of HsdM could be that the central conserved region loosely associates with a second M subunit. Taking the errors into account though, the CCHalfSHis MTase could have also purified in an even ratio. Nevertheless, in both cases the purifications seemed to be providing enough of the proteins to form active MTases.

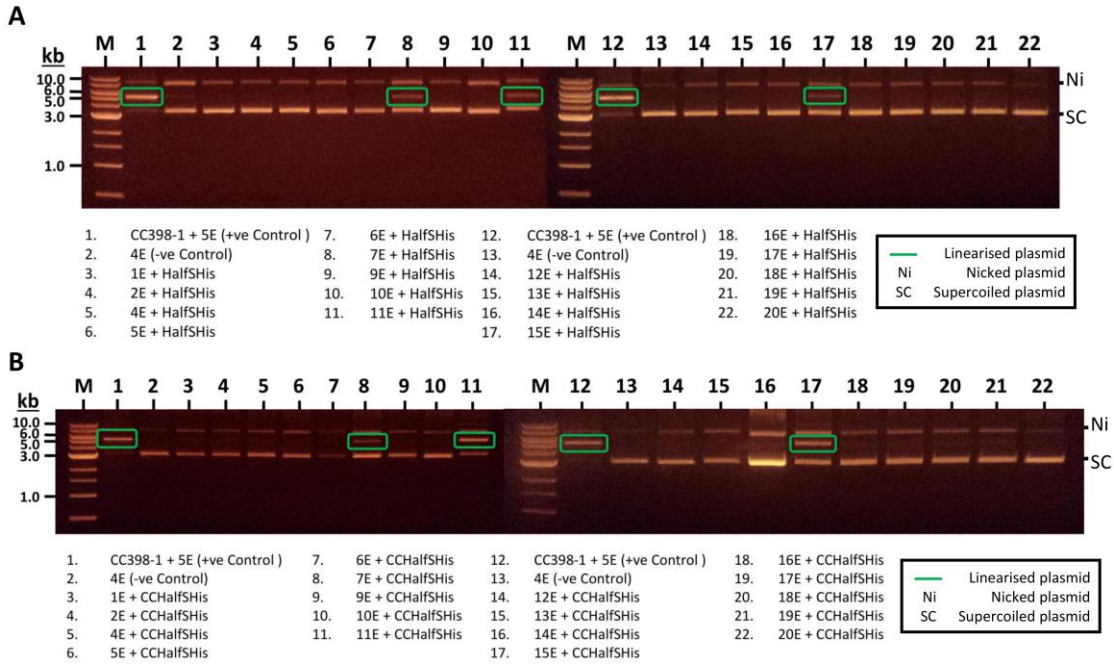


Figure 97: Plasmid cleavage assay of HalfSHis (A) and CCHalfSHis (B).

Having successfully purified the two half HsdS proteins, they could be assayed for *in vitro* activity. They were used in plasmid cleavage assays against the Eddy library. At this stage in the investigation, it had already been shown that the truncated genes produced viable MTases, but also their DNA sequence specificity had been established. It was interesting to see whether

the pattern of plasmid cleavage was in agreement with the pattern that was expected from the results from SMRT sequencing.

The linearised 7E, 11E and 15E plasmid species from the plasmid cleavage assay showed that the HalfSHis MTase was active *in vitro*, and that it possessed the expected DNA recognition sequence of $ACCN_5GGT$ (interpreted by RMSearch) (Fig. 97A). This was also the case for the CCHalfSHis MTase (Fig. 97B). Given this success, the activity of the two MTases was checked against modified and unmodified genomic DNA (Fig. 98).

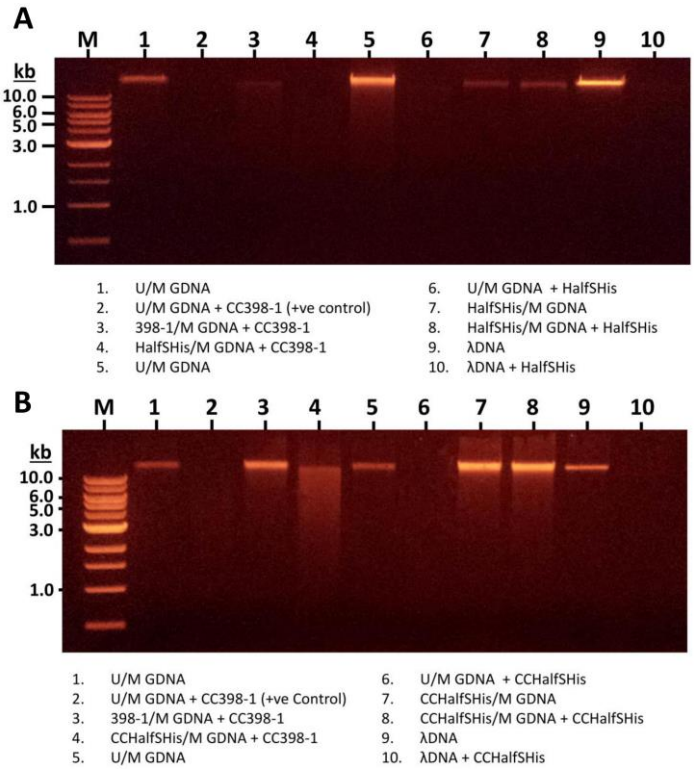


Figure 98: Genomic DNA (GDNA) restriction assay of HalfSHis (A) and CCHalfSHis (B). On both gels, Lane 1 contains unmodified (U/M) genomic DNA alone. Lane 2 shows that with the addition of wild-type CC398-1 MTase (+ve control), the genomic DNA no longer appears on the gel. 398-1/M GDNA denotes genomic DNA that has been modified by the wild-type enzyme. HalfSHis/M denotes genomic DNA that has been modified by the HalfSHis protein. CCHalfSHis/M denotes genomic DNA that has been modified by the CCHalfSHis protein. On both gels, Lanes 9 and 10 contain lambda (λ) phage DNA.

Both MTases showed *in vitro* restriction of phage DNA (Lane 10 A and B), and unmodified *E. coli* genomic DNA (Lane 6, A and B). As hoped, both MTases were inactive against self-modified genomic DNA (Lane 7 and 8, A and B).

The half HsdS enzymes had activity *in vivo* and *in vitro*. The data obtained from these experiments had indicated that the half HsdS enzymes were dimerising, in order to become active. Gel filtration HPLC was used to gain evidence for this arrangement in their quaternary structure (Fig. 99).

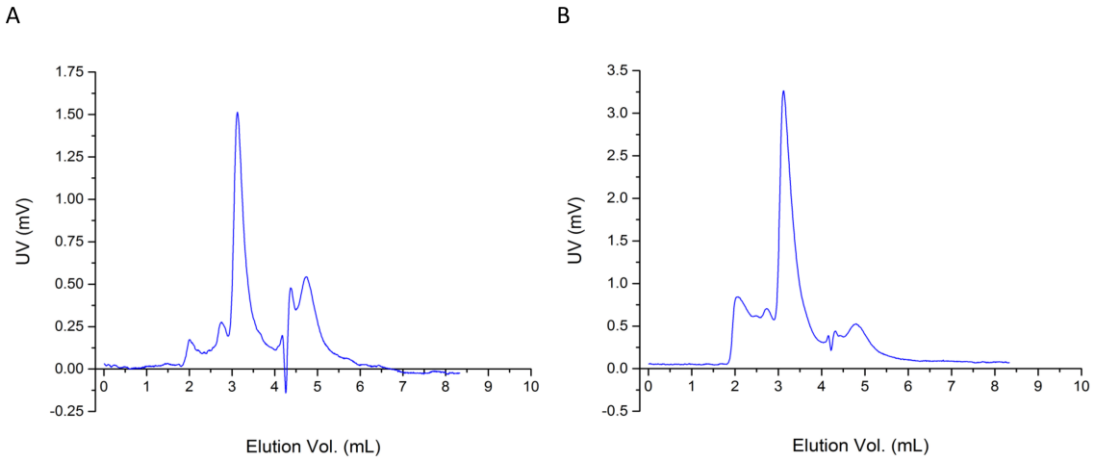


Figure 99: Gel filtration HPLC elution profile of the HalfSHis (A) and CCHalfSHis (B) MTases.

The elution profiles of the two MTases were similar. Both showed clean, sharp peaks, corresponding to the target enzymes, although both also contained contaminating species. Due to the small difference in molecular weight of the two half hsdS species, it was not possible to resolve the difference on the HPLC, so both had an elution volume of 3.125 mL. However, as both gave a retention time similar to the wild-type enzyme (3.1 mL), it can be assumed that this is indeed further evidence for the dimerisation of the half HsdS enzymes.

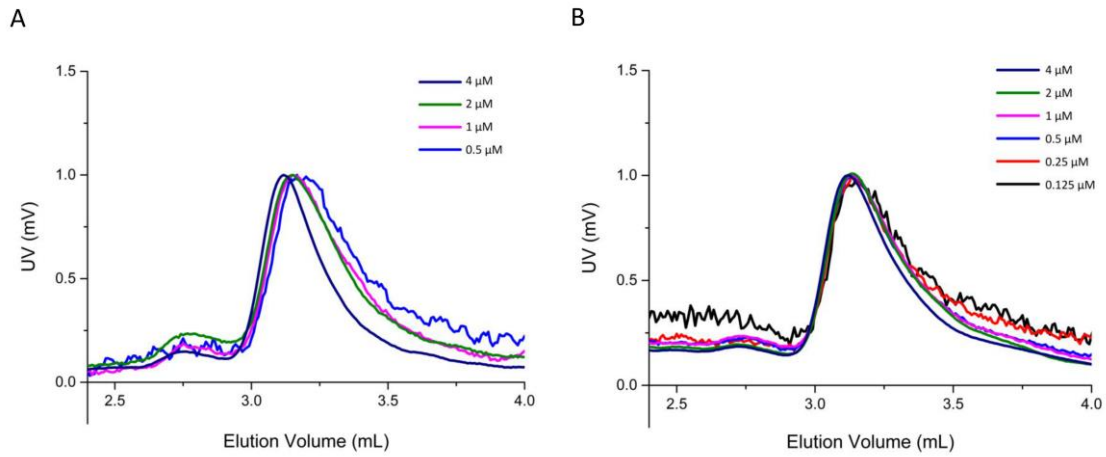


Figure 100: Normalised elution profiles of decreasing concentrations of the HalfS MTase (A) and the CCHalfS MTase (B).

Comparing the elution profiles of decreasing concentrations of the half HsdS enzymes gave interesting results (Fig. 100) The break-down of the HalfSHis enzyme appears to occur at a concentration of 2 μM, as the elution profiles shift to the right. This is far higher than any of the previous MTases. What was encouraging though, was that the CCHalfSHis appeared to be far more stable, and only began to turn to its M_1S_1 form at a concentration of 0.25 μM. This supports the idea that the addition of central conserved region is allowing the CCHalfSHis to form the wild-type, circular configuration, but also bind the second HsdM more strongly.

3.7. CC398-1 HsdM to Half HsdS Fusion

With the structure of the Type IIG systems in mind, the natural progression from the MS fusion and half HsdS enzymes was to create an M to HalfS fusion. Results indicated that both half HsdS MTases were active, and so it was not hugely important which of these was used as a template. At the time of manufacture however, it was considered that retaining the entire central conserved region would more likely produce an active enzyme. Therefore, the primers used to create the CCHalfS MTase were used, with MS fusion gene as a template. This produced an *hsdM* to half *hsdS* gene fusion, which incorporated all of the central conserved region. This was called MCCHalfS fusion or MCHSF.

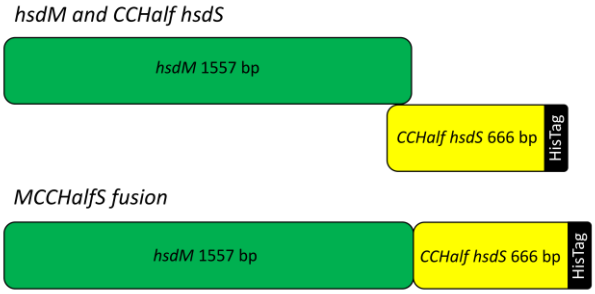
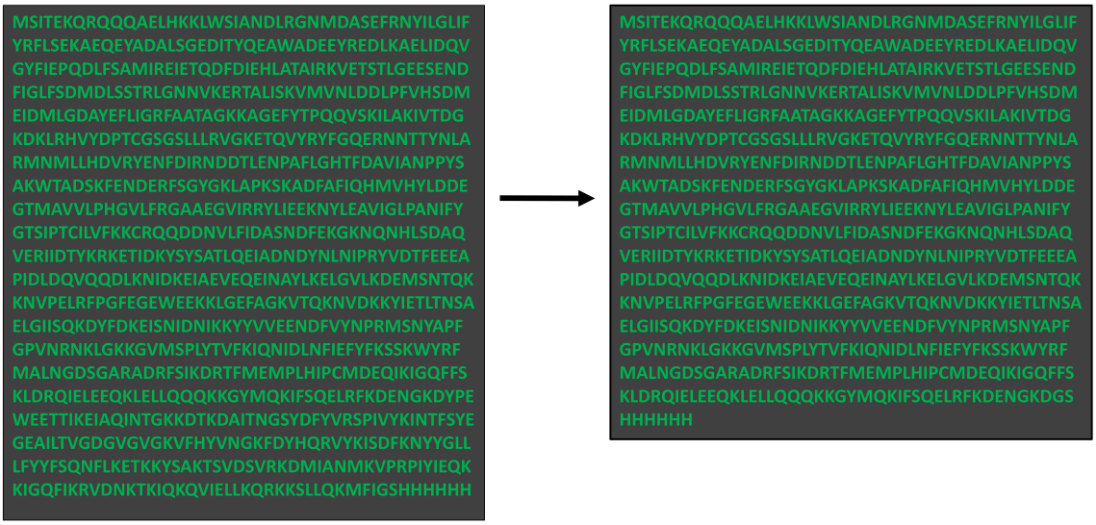


Figure 101: Cartoon diagram of the genetic rearrangement of CCHalfS to MCCHalfS fusion.

The new MCCHalfS fusion protein would have the same amino acid sequence as the CCHalfS MTase but would be a single peptide (Fig. 102A)

A



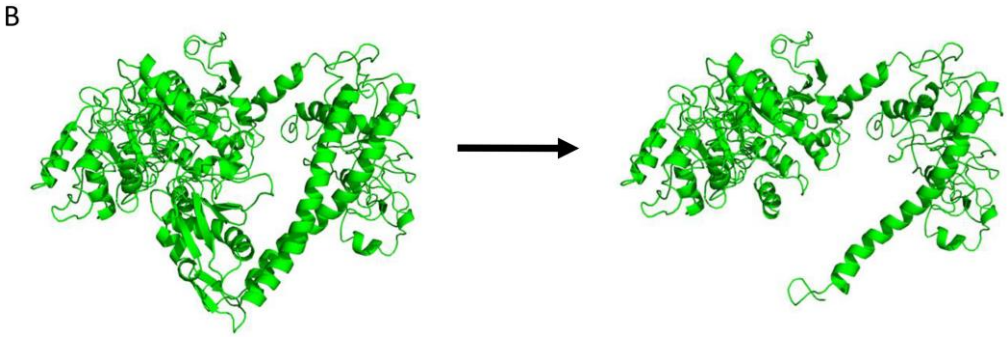


Figure 102: Amino acid sequence of the MS fusion, truncated to the MCCHalfS fusion (A) and the corresponding structural change on the protein model (B).

A PCR product of the expected size for the MCCHalfS fusion gene was purified and ligated into vector pJF (see Appendix C for plasmid map), the product of which was sequenced and verified.

The MCCHalfS fusion was expressed on a small scale and the solubility of the protein product was confirmed. The protein was then produced on a larger scale and purified by nickel affinity and size exclusion chromatography. Selected samples from these purification steps were analysed by SDS-PAGE (Fig. 103).

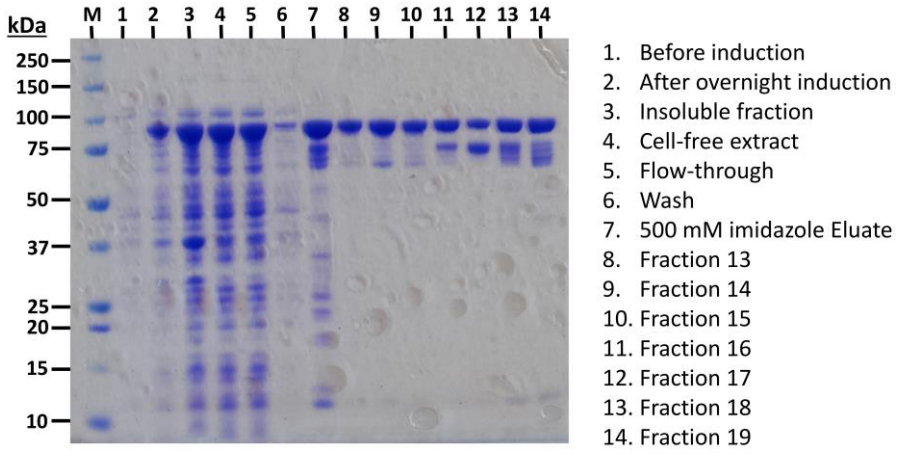


Figure 103: SDS-PAGE gel of the nickel affinity and size exclusion chromatography purification of MCCHalfS protein.

The MCCHalfS fusion had expressed successfully, and the protein appeared to purify well. UV/vis spectroscopy was used to estimate the concentration of the protein, which was then assayed for restriction activity.

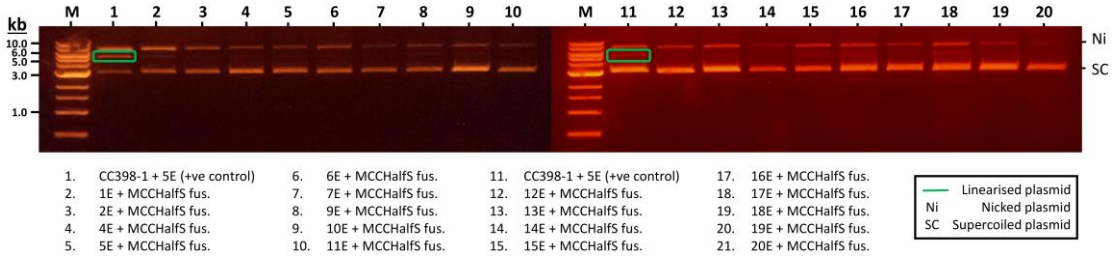


Figure 104: Agarose gel electrophoresis of the results of a plasmid cleavage assay of the MCCHalfS fusion protein.

A lack of a coherent cutting pattern through several plasmid cleavage assays showed that this fusion preparation did not possess restriction activity (Fig. 104). Therefore, the *in vivo* assay was used as a quick way of identifying whether the new construct possessed activity. This would establish whether it was worthwhile to pursue investigations *in vitro*.

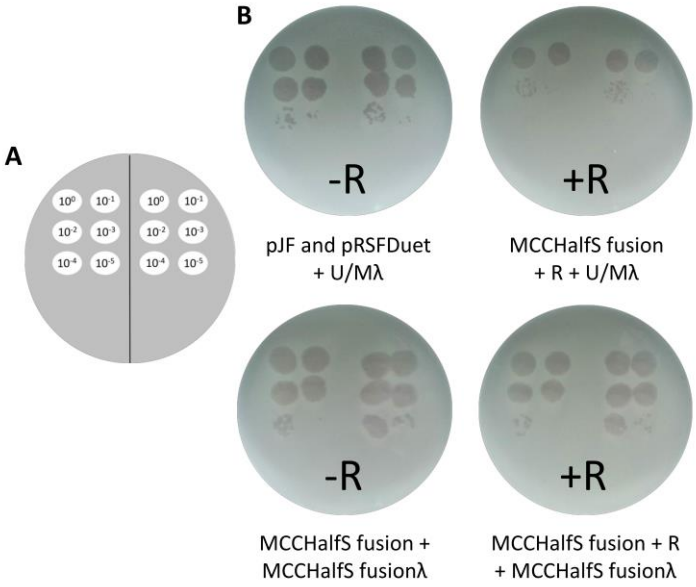


Figure 105: Diagram of dilutions (A) *In vivo* spot tests to assay the MCCHalfS fusion for R-M activity (B).

In vivo spot tests showed that the MCCHalfS fusion was restriction and modification active (Fig. 105). When tested against unmodified (U/M) λ phage, there was cut back by two log dilutions compared to the control. When λ phage that had been used to infect cells containing the MCCHalfS fusion gene was used to infect cells containing the restriction active fusion, there was no difference in plaque formation compared to the control. This indicated that the λ phage had been modified by the MCCHalfS fusion and was subsequently resistant to restriction by the restriction active form of the fusion. These results were verified by conducting several full plate *in vivo* assays of this MTase.

MCCHalfS fusion MTase *In vivo* Assay Results:

Phage Type	Phage Dilution	Phage Volume	R-M System	Number of Plaques
Unmodifiedλ	10^{-6}	100 µL	None	43
	10^{-5}	100 µL	MCCHalfS fus. + R	57
	10^{-6}	100 µL	None	19
	10^{-5}	100 µL	MCCHalfS fus. + R	8
	10^{-6}	100 µL	None	51
	10^{-5}	100 µL	MCCHalfS fus. + R	46
MCCHalfS fus.λ	10^{-6}	100 µL	MCCHalfS fus.	87
	10^{-6}	100 µL	MCCHalfS fus. + R	98
	10^{-6}	100 µL	MCCHalfS fus.	133
	10^{-6}	100 µL	MCCHalfS fus. + R	142
	10^{-6}	100 µL	MCCHalfS fus.	145
	10^{-6}	100 µL	MCCHalfS fus. + R	133

Table 6: Raw data from full plate *in vivo* assays of the MCCHalfS fusion MTase. The table shows ±R pairs, which are the results from the experiment (+R) and control (-R), and the subsequent repeats. The plaque numbers cannot be compared without adjusting for volume and dilution.

The E.O.P. of the unmodified λ phage against the restriction active MCCHalfS fusion was 0.10 ±0.04 (see Appendix E for calculations). The MCCHalfS fusion protein appeared to be forming a highly active restriction complex. This value was the same as that calculated for the MS fusion, and suggests the possibility that the subunits being contained in a single peptide was having a positive effect on activity. It was hard to compare the MTases accurately given the high error values, but it did confirm that the MCCHalfS fusion is restriction active. The MCCHalfS fusion modified λ phage produced an E.O.P. of 1.02 ±0.26 (see Appendix E for calculations). This result is higher than that of the wild-type enzyme, but again is within error. It does suggest that the MCCHalfS fusion MTase is a highly efficient methylase. A possible explanation for this is its ability to form the wild-type, circular conformation, due to the inclusion of all of the central conserved region.

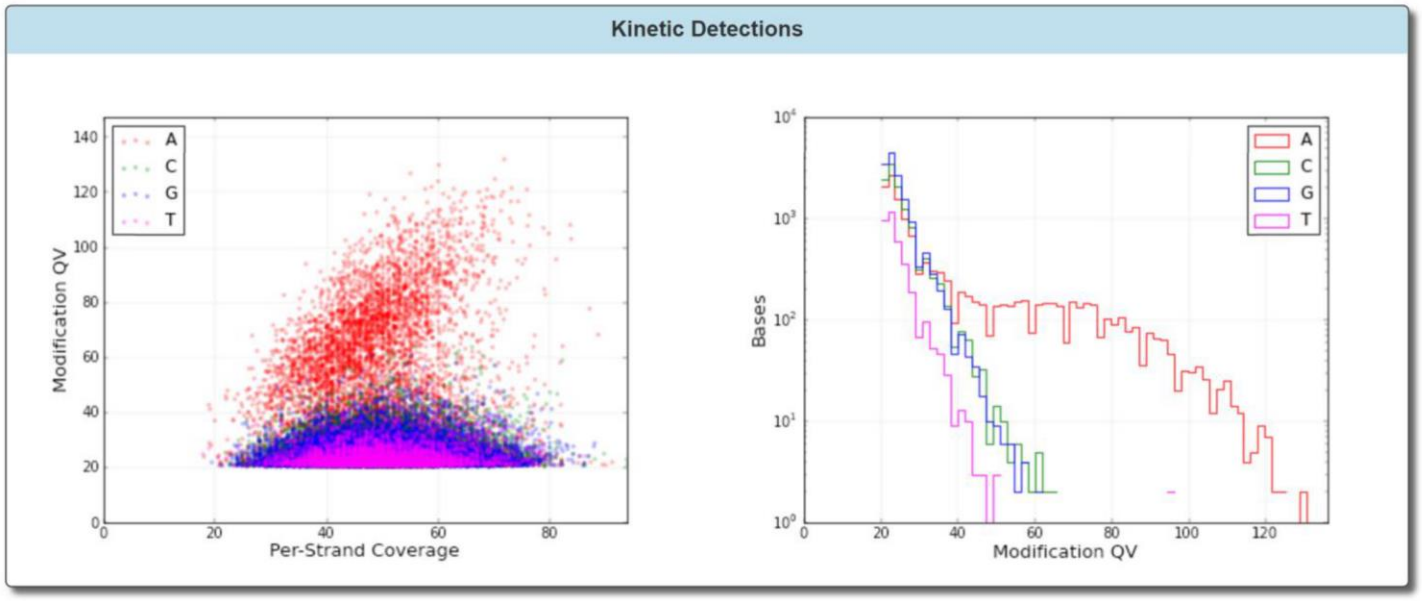
Having observed activity *in vivo,* SMRT sequencing was then used to identify any pattern of methylation that the MCCHalfS fusion might cause. *E. coli* genomic DNA, modified by the MCCHalfS fusion was SMRT sequenced and the results were analysed by *Pacific Biosciences* software (Fig 106).

MCCHalfS fusion MTase SMRT Sequencing Results:

A **Reports for Job Bowers_MCHSF_Mods**

PACIFIC BIOSCIENCES

SMRT Cells: 2 Movies: 2



B **Reports for Job Bowers_MCHSF_Mods**

PACIFIC BIOSCIENCES

SMRT Cells: 2 Movies: 2

Motif Summary

Motifs	Modified Position	Type	% Motifs Detected	# Of Motifs Detected	# Of Motifs In Genome	Mean Modification QV	Mean Motif Coverage	Partner Motif
ACCNNNNNGGT	1	m6A	99.66%	2655	2664	73.31	47.30	ACCNNNNNGGT
BHACCNNNNNGGV	3	m6A	32.29%	1355	4196	45.68	52.42	
YAHACCNNNNNGGV	4	m6A	26.96%	151	560	43.58	55.44	
BNATCNNNNNGGTGG	3	m6A	21.2%	39	184	47.97	47.87	CCACCNNNNNGATNV
CCACCNNNNNGATNV	3	m6A	18.48%	34	184	46.56	48.79	BNATCNNNNNGGTGG
GACCNNNNVGGVR	2	m6A	21.0%	109	519	41.97	51.04	
AANACCNNNNSGGV	4	m6A	15.97%	42	263	41.88	49.83	

Figure 106: Results from SMRT sequencing (*Pacific Biosciences*) of MCCHalfS fusion modified *E. coli* genomic DNA. Kinetic Detections (A). The first graph is a scatter plot showing the detection of base methylation. The higher coverage and detection of adenine methylation (red dots) confirms the MCCHalfS fusion protein forms an adenine methylase. The second graph shows detection of bases against detection of methylated bases. This too confirms adenine methylation. Table to show statistical data for the modified motifs (B). The table shows the single motif that was detected, $ACCN_5GGT$.

Results from the half HsdS MTases gave an indication of the mode of action of the MCCHalfS fusion. SMRT sequencing confirmed that the new fusion did indeed share the same DNA recognition sequence as the half HsdS MTases ($ACCN_5GGT$) (see Appendix G for additional results). This palindromic sequence was further evidence that the MCCHalfS fusion was dimerising.

It was considered possible that the protein form of this new fusion suffered from similar issues to the half HsdS MTases, and as such would need to be purified using a faster method. The Quick purification method was used to prepare more of the protein, the presence and quality of which was then checked by SDS-PAGE (Fig. 107).

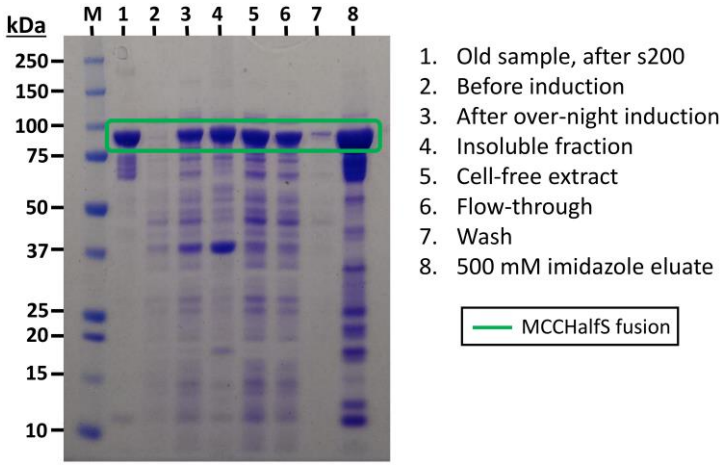


Figure 107: SDS-PAGE analysis of samples from the "Quick" purification of MCCHalfS fusion protein.

SDS-PAGE analysis confirmed that the Quick purification method had successfully retained and purified the MCCHalfS fusion protein. In Lane 1 of the gel was run a sample of the same protein from the previous preparation. It was quite apparent that the previous method produced a far cleaner sample, but one that showed no signs of restriction activity. The new MCCHalfS fusion protein sample alone was analysed by SDS-PAGE, and showed few contaminants (Fig. 108). It was hoped that activity would be observed now that the speed of the process had increased. The level of contamination in the sample made it difficult to get an accurate concentration for the target protein. However, the yield was estimated to be 11.3 mg from a 5 g cell pellet.

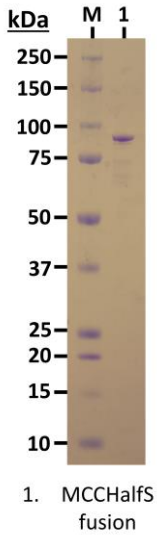


Figure 108: SDS-PAGE gel of the purified MCCHalfS fusion protein.

In theory, this half S fusion was dimerising. This idea was supported by data from SMRT sequencing, but further evidence was sought via gel filtration HPLC (Fig. 109).

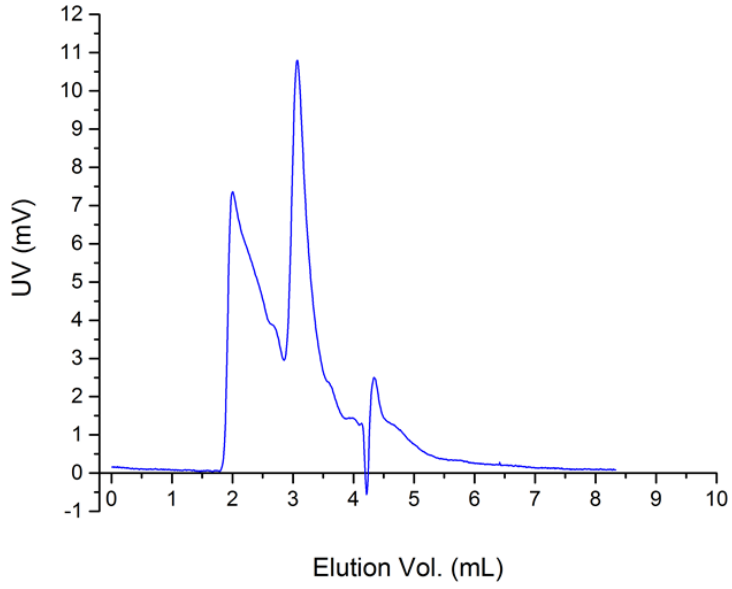


Figure 109: Gel filtration HPLC elution profile of the MCCHalfS fusion protein, after Quick purification.

The elution profile of the protein was untidy. A large peak at the start of the trace corresponded to a very high molecular weight. This peak had noticeable shoulders, indicating aggregated species, which was eluting into the void volume. The largest peak on the trace was assumed to have been created by the target protein. Once again, the resolution of this protein was poor

but a similar elution volume (3.0 to 3.2 mL) to the wild-type enzyme indicated that the MCCHalfS fusion was forming a dimer in solution.

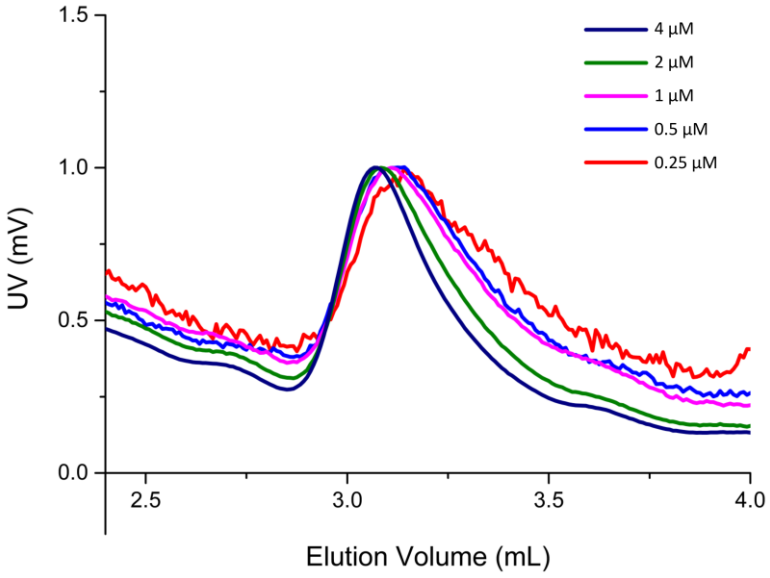


Figure 110: Normalised elution profiles of decreasing concentrations of the MCCHalfS fusion MTase.

Once again, the elution profiles of decreasing concentrations of the protein were compared (Fig. 110). This MTase appeared to have a similar level of stability to the HalfSHis enzyme. In both cases the elution profiles begin to shift to the right at a concentration of 2 µM. However, with this enzyme, the HsdM and half HsdS are fused and so this shift would correspond to the dissociation of two $M_1S_{½}$ fusion monomers. Gel filtration HPLC results from the CCHalfSHis MTase suggested that the inclusion of the central conserved region was having a positive effect on the stability of the enzyme complex. Despite the presence of this same peptide sequence in this protein, this fusion appears relatively unstable. When purified, this MTase would often precipitate and/ or lose activity within a few days, despite storage in 50 % glycerol at -20 °C. The lack of stability seen in HPLC was most likely due to the fragility of the fusion peptide itself, and not its ability to form a dimer.

The fusion protein had purified successfully and seemed to be adopting a conformation that would generate activity in solution. The plasmid cleavage assay was used to identify *in vitro* restriction activity (Fig. 111).

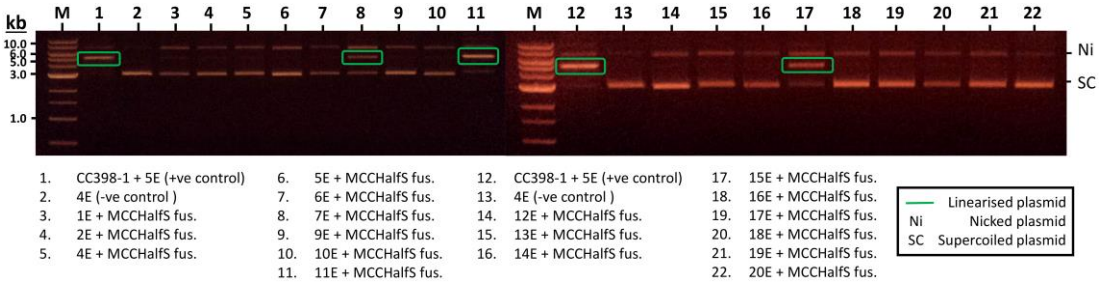


Figure 111: Agarose gel electrophoresis of samples from a plasmid cleavage assay of the MCCHalfS fusion protein.

The pattern of cleavage in the plasmid cleavage assay of the new preparation of the MCCHalfS fusion was identical to that of the half HsdS MTases (7E, 11E and 15E positive). This was an exciting result, confirming that the MCCHalfS fusion was active *in vitro.*

A genomic DNA cleavage assay of the MCCHalfS fusion MTase showed that the fusion was also restriction active against genomic DNA and λ phage DNA (Fig. 112).

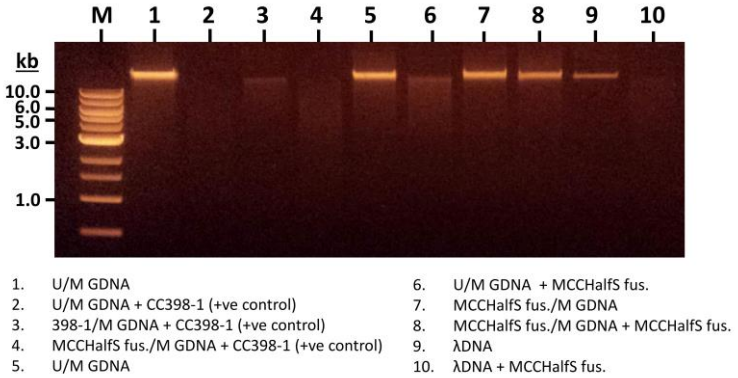


Figure 112: Genomic DNA (GDNA) cleavage assay of the MCCHalfS fusion MTase. Lane 1 contains unmodified (U/M) genomic DNA alone. Lane 2 shows that with the addition of wild-type CC398-1 MTase (+ve control), the genomic DNA no longer appears on the gel. 398-1/M GDNA denotes genomic DNA that has been modified by the wild-type enzyme. MCCHalfS fus./M denotes genomic DNA that has been modified by the MCCHalfS fusion protein. On both gels, Lanes 9 and 10 contain lambda (λ) phage DNA.

Chapter Four:

Discussion and Conclusions

The aim of this work was to show the structural similarities between Type I and Type II R-M systems, to provide evidence for their evolutionary relationship. It sought to manipulate the structure of a Type I R-M system, to create a structure similar to a Type II system. The Type II systems are categorised into several different sub-types. The Type IIB systems cleave double-stranded DNA either side of their recognition sequence, whilst Type IIG systems are a single peptide, and so cleave at only one side. Both sub-types carry out both restriction and methylation functions, and are dependent on S-adenosylmethionine. This project intended to engineer an enzyme with these properties, using a Sau1 Type I system as a template.

Removing the motor domain from the CC5 HsdR, and fusing it to the CC398-1 MTase, attempted to make a complete R-M system in a single peptide. Type I systems bind their recognition sequence but then translocate the DNA at both ends, eventually cleaving it at unpredictable locations either side of their recognition sequence. Removing the motor domain would prevent the movement of the DNA and elicit the same type of action as the Type IIB enzymes. By fusing HsdR to M, the R-M domains would be brought into a single peptide, which would be SAM dependent. This would create a Type IIB-like structure (Fig 113).

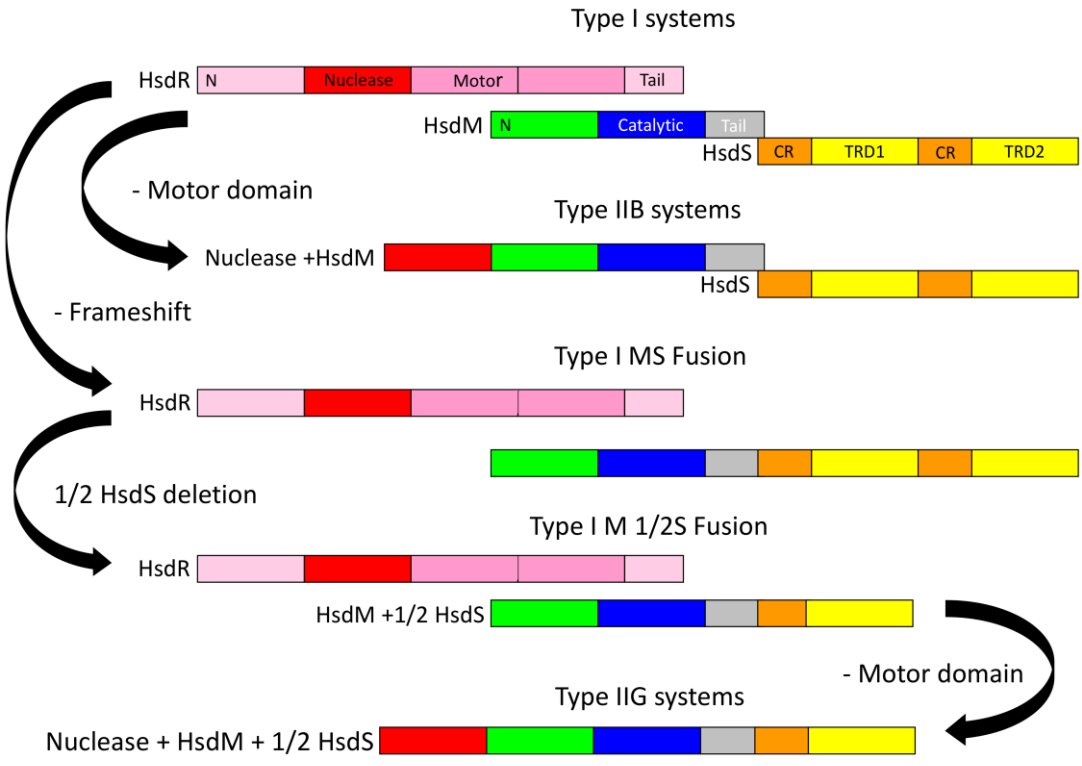


Figure 113: A step-wise process, via which changes to Type I systems can form Type IIB and IIG systems.

As there is still very little structural information on the Sau1 enzymes, it was extremely difficult to estimate the amino acid sequence of the CC5 HsdR that was specific only to motor activity. Structural data shows that the HsdR contains four distinct domains, the first of which is the nuclease (Lapkouski et al. 2009). However, it has been shown that mutations to one of the domains can have an effect on the activity of another (Šišáková, Weiserová, et al. 2008). It is therefore unlikely to be as simple as removing all the sequence after the nuclease, in the hope of retaining cleavage activity. Given this problem, a high-throughput method of engineering and subsequent screening would be necessary to identify a restriction-active construct. The method used to create the gene fusions was not reliable and the method of screening, involving large-scale expression and purification, was time consuming. *In vivo* assays were a quicker way of conducting this process, but unfortunately returned no restriction-active candidates. What is encouraging however, is that all the R-M fusions studied did seem to possess methylase activity. This suggests that the process can create an enzyme, which can fold correctly and produce activity but that it was the HsdR domain that was not viable. All it would take therefore, is further attempts with several different truncations to the R subunit. On the other hand, perhaps simply truncating the HsdR sequence is an error. If all the domains help coordinate each activity, then removing any part may prevent the desired nuclease activity. Additionally, structural data shows that the C-terminal, helical domain, interacts with the MTase complex. This suggests that a new approach should focus on deleting the two central motor domains, or mutating them to inhibit their activity. The primary target for this approach should be the DEAD-box motifs, which are necessary for activity on all helicases.

A great deal more success was found when concentrating on the Type I MTases. This project shows the incremental change in structure, from the wild-type CC398-1 Type I MTase, to an active MTase in a single, fused peptide, to a single peptide MTase with one TRD. The characteristics of this final, novel construct are shared by the Type IIG system, BpuSI. Although this construct produced an apparently fragile protein, it did possess activity both *in vivo* and *in vitro*. The structure dimerises, and therefore recognises a symmetrical DNA sequence. This aspect brings the structural relationship closer to the Type IIB systems, which also dimerise. The IIB systems do however possess two TRDs, and so a better comparison would be the MS fusion. Unfortunately, the MS fusion created here did not dimerise but took on a pseudo M_2S_1 conformation, due to the presence of a bound product of proteolysis. Despite attempts at further purification, this extra fragment remained bound to the MS fusion. More

stringent methods could be used to remove this contaminant, in order to investigate the quaternary structure of the protein more thoroughly.

Size exclusion HPLC results from the CCHalfSHis enzyme support the theory that the central conserved region is important to the domain organisation of these MTases. However, evidence of activity *in vivo* and *in vitro* with the HalfSHis enzyme showed that the complete central conserved region does not appear entirely necessary. If one were to imagine the proposed circular organisation of the M and S subunits, the two TRDs are bridged by the central conserved region and the sequences of N and C-termini. In addition, the central conserved region is proposed to correspond to the N number of the recognition sequence. In the case of the two HalfS species presented here, the first TRD possesses both the central conserved region (only in part, in the case of HalfSHis) and the N-terminal sequence, which is more than half of the total peptide sequence of the HsdS. Upon direct, homogenous dimerisation of these HalfS species, it is possible that the circular structure is larger than wild-type. It is therefore surprising that both species recognise the same sequence, it is shorter than that of the wild-type sequence, and that the methylated bases are further apart. In the case of the TRD-2 deletion mutant created by Abadjieva *et al*., the new palindromic sequence was the same length as that of the wild-type and the distance between methylated bases remained the same (GA<u>A</u>N_7<u>T</u>TC and GA<u>A</u>N_6R<u>T</u>CG respectively). This mutant also possessed the N-terminal sequence and the entire central conserved region sequence (Abadjieva et al. 1993; Abadjieva et al. 2003). Taylor et al., 1994 and Janscak & Bickle, 1998 showed that alternative C-termini can be introduced to the HsdS, but that it would maintain MTase activity. Here, it is shown that the addition of amino acid sequence does not affect activity, nor is it affected by the lack of a complete central conserved region. Size exclusion HPLC data indicate that the second HsdM of the HalfSHis complex dissociates at a higher concentration, relative to most of the other MTases examined here. A reasonable assumption for why this is happening is the missing section of the central conserved region leaves less of the HsdS sequence to which the HsdM can bind, and as such it has less affinity for it. The HPLC data, SMRT sequencing results and other evidence of activity suggest that the HalfSHis does still form an active dimer, and therefore is adopting the circular conformation. Unfortunately, these proteins appear relatively unstable, making the acquisition of structural information by x-ray crystallography unlikely. Therefore, it will be difficult to discover how the two HalfS monomers are interacting. However, these results do indicate that the central conserved region is not acting in the manner in which it has hitherto been thought. The amino acid sequence between the two

TRDs cannot be directly specific to the random sequence between the two parts of the bipartite recognition sequence.

When investigating the relationship between Types I and II, Type ISP systems are clearly important. Given that they possess both modification and restriction activities in a single peptide, they can be seen as the link that bridges the gap between Type I systems and Type IIG systems. The proposal by Kennaway *et al.* (2012) was that IIG enzymes should be considered half of a Type I enzyme, which is lacking its motor domains. In the Type ISP systems, we have a structure akin to a Type IIG enzyme *with* an ATP-dependent helicase domain (Fig. 114).

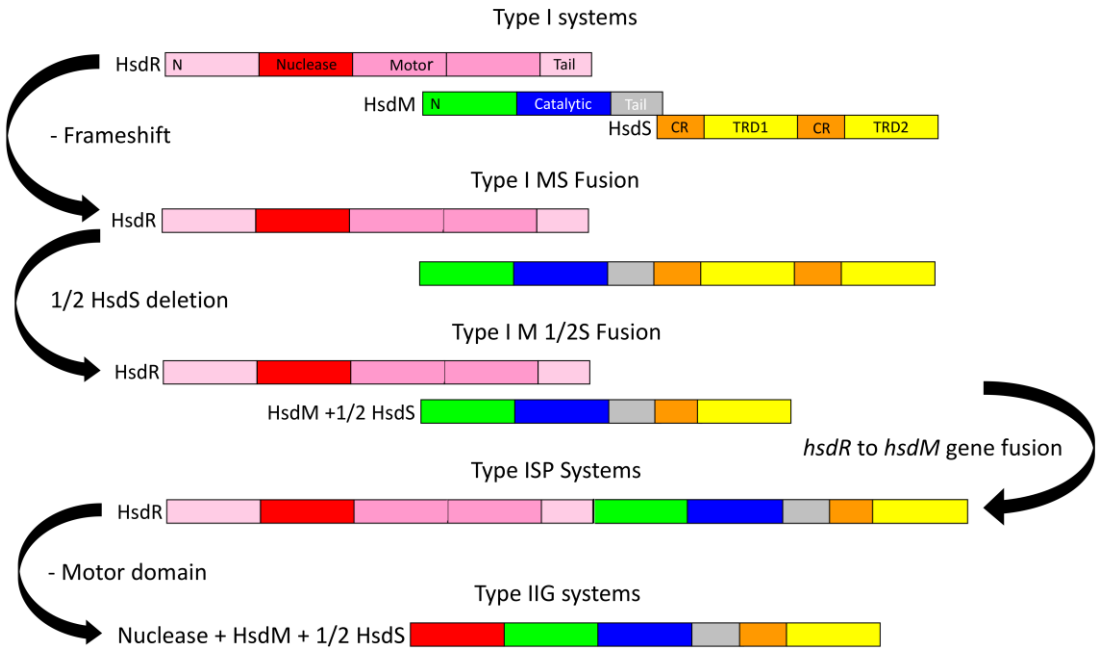


Figure 114: Proposed progression from Type I systems to Type IIG, via the Type ISP link.

The creation of an MTase and REase active MCCHalfS fusion is the final step before a Type ISP system (Fig. 114). This suggests that the next step in this investigation is not to remove the motor domains from the HsdR at all, but to fuse the entire subunit to the N-terminal of the MCCHalfS fusion. This would give it a structure similar to a Type ISP. The step after this would be to remove some or all of the peptide sequence from the motor domains, in an effort to make the final step to a Type IIG system.

A key aim in this project was to create a new restriction-active enzyme, which had a structure similar to that of a Type II R-M system. Although this was unsuccessful, this work has

identified that a significant portion of an HsdR can be appended to an MTase, without loss of methylation activity. It has also highlighted that this task should be approached differently, with an emphasis on decreasing the time taken to rule out inactive constructs. Fortunately, there were several other aspects of the project that proved successful and therefore indicate a method that could be applied to achieve the main goal of producing a DNA-cutting enzyme. By mimicking the step by step changes proposed for the evolution of these systems, viable structural changes were engineered incrementally. A new M to S fusion was created, and two novel half S subunits were created, purified and assayed *in vitro*. This had not before been achieved. The M to half S fusion is entirely new and was also successfully purified and assayed *in vitro*. Not only that, but these modifications elicited a new topology and a new DNA recognition sequence. This work satisfies its goal to support the claim that Types I and II are evolutionarily linked. It has shown that these structures can be manipulated to produce different viable enzymes, indicating that this could have occurred through an evolutionary process.

RM fusion Results Summary:

Construct Name	Gene Diagram	Protein Purification	*In vivo* Restriction	*In vivo* Modification
RM_CM_1		−	N/A	N/A
RM_CM_2		+	−	+
RM_EB_1		−	N/A	N/A
RM_EB_2		+	−	+
RM_EB_3		−	−	+
RM_EB_4		−	−	+
RM_EB_5		−	−	+
RM_EB_6		−	−	+
RM_EB_7		−	−	+
RM_EB_8		−	−	+
RM_EB_9		−	−	+
RM_EB_10		−	−	+
RM_EB_11		−	−	+

Table 7: A summary of the results from the work to create a Type I R to M fusion. It was subsequently discovered that the lack of results from RM_CM_1 and RM_EB_1 was due to inviable gene sequences. Results for these two constructs were therefore deemed not applicable (N/A).

MTase Results Summary:

MTase	Gene Diagram	EOP Unmodified λ phage	EOP Modified λ phage	SMRT Target	Protein Purification	Eddy Plasmid(s) Cleaved	MW (kDa)	MW (SEC HPLC) (kDa)
CC398-1 (wild-type)		0.15 ±75.0 %	0.88 ±30.6 %	ACCNNNNNRTGA	✚	5E 6E 7E 12E 14E	166 (M_2S_1)	210
MS fusion		0.09 ±52.3 %	0.78 ±17.6 %	ACCNNNNNRTGA	✚	5E 6E 7E 12E 14E	106 (monomer)	311.5
HalfSHis		0.40 ±49.9 %	0.70 ±28.5 %	ACCNNNNNGGT	✚	7E 11E 15E	170.6 (dimer)	199.6
CCHalfSHis		0.58 ±36.9 %	0.75 ±20.1 %	ACCNNNNNGGT	✚	7E 11E 15E	172.8 (dimer)	199.6
HalfS (without HisTag)		N/A	N/A	N/A	▬	N/A	168.7 (dimer)	N/A
CCHalfS (without HisTag)		N/A	N/A	N/A	▬	N/A	170.9 (dimer)	N/A
MCCHalfS fusion		0.1 ±36.4 %	1.02 ±25.6 %	ACCNNNNNGGT	✚	7E 11E 15E	172.8 (dimer)	242.5

Table 8: Summary of the results from the work on the CC398-1 MTase.

References

Abadjieva, A., Patel, J., Webb, M., Zinkevich, V., Firman, K., 1993. A deletion mutant of the type IC restriction endonuclease EcoR1241 expressing a novel DNA specificity. *Nucleic Acids Research*, 21(19), pp.4435–4443.

Abelson, J., Trotta, C.R. & Li, H., 1998. tRNA Splicing*. *The Journal of Biological Chemistry*, 273(21), pp.12685–12689.

Adamczyk-Poplawska, M., Lower, M. & Piekarowicz, A., 2011. Deletion of one nucleotide within the homonucleotide tract present in the hsdS gene alters the DNA sequence specificity of type I restriction-modification system NgoAV. *Journal of Bacteriology*, 193(23), pp.6750–9.

Ain, Q.U., Chung, J.Y. & Kim, Y.-H., 2015. Current and future delivery systems for engineered nucleases: ZFN, TALEN and RGEN. *Journal of Controlled Release*, 205(10), pp.120–127.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), pp.403–410.

Arber, W. & Dussoix, D., 1962. Host specificity of DNA produced by Escherichia coli. *Journal of Molecular Biology*, 5(1), pp.18–36.

Armand-Lefevre, L., Ruimy, R. & Andremont, A., 2005. Clonal Comparison of Staphylococcus aureus Isolates from Healthy Pig Farmers, Human Controls, and Pigs. *Emerging Infectious Diseases*, 11(5), pp.711–714.

Atanasiu, C., Byron, O., McMiken, H., Sturrock, S.S., Dryden, D.T.F., 2001. Characterisation of the structure of ocr, the gene 0.3 protein of bacteriophage T7. *Nucleic Acids Research*, 29(14), pp.3059–3068.

Ballhausen, B., Kriegeskorte, A., van Alen, S., Jung, P., Köck, R., Peters, G., Bischoff, M., Becker, K., 2016. The pathogenicity and host adaptation of livestock-associated MRSA CC398. *Veterinary Microbiology*.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. *Nucleic Acids Research* , 28(1), pp.235–242.

Bheemanaik, S., Reddy, Y.V.R. & Rao, D.N., 2006. Structure, function and mechanism of exocyclic DNA methyltransferases. *Journal of Biochemistry*, 190(399), pp.177–190.

Bickle, T.A. & Kruger, D.H., 1993. Biology of DNA Restriction. *Microbiological Reviews*, 57(2), pp.434–450.

Bird, A., 2007. Perceptions of epigenetics. *Nature*, 447(7143), pp.396–398.

Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A., Bonas, U., 2009. Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors. *Science*, 326(5959), pp.1509–1512.

Brumfitt, W. & Hamilton-Miller, J., 1989. Methicillin-Resistant Staphylococcus aureus. *New England Journal of Medicine*, 320(18), pp.1188–1196.

Bujnicki, J.M., 2001. Understanding the evolution of restriction-modification sys- tems : Clues from sequence and structure comparisons . *Acta Biochimica Polonica*, 48(4), pp.935–67.

Carroll, D., 2011. Zinc-Finger Nucleases: A Panoramic View. *Current Gene Therapy*, 11(1), pp.2–10.

Casjens, S.R. & Hendrix, R.W., 2015. Bacteriophage lambda: Early pioneer and still relevant. *Virology*, 479–480, pp.310–330.

Ceasar, S.A., Rajan, V., Prykhozij, S.V., Berman, J.N., Ignacimuthu, S., 2016. Insert, remove or replace: A highly advanced genome editing system using CRISPR/Cas9. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1863(9), pp.2333–2344.

Champoux, J.J., 2001. DNA TOPOISOMERASES: structure, function, and mechanism. *Annual Review of Biochemistry*, 70, pp.369–413.

Chand, M.K., Nirwan, N., Diffin, F.M., van Aeist, K., Kulkarni, M., Pernstich, C., Szczelkun, M.D., Saikrishnan, K., 2015. Translocation-coupled DNA cleavage by the Type ISP restriction-modification enzymes. *Nat Chem Biol*, 11(11), pp.870–877.

Chandrasegaran, S. & Carroll, D., 2016. Origins of Programmable Nucleases for Genome Engineering. *Journal of Molecular Biology*, 428(5, Part B), pp.963–989.

Chellat, M.F., Raguž, L. & Riedl, R., 2016. Targeting Antibiotic Resistance. *Angewandte Chemie International Edition*, 55(23), pp.6600–6626.

Chen, K., Roberts, G.A., Stephanou, A.S., Cooper, L.P., White, J.H., Dryden, D.T., 2010. Fusion of GFP to the M.EcoKI DNA methyltransferase produces a new probe of Type I DNA restriction and modification enzymes. *Biochemical and Biophysical Research Communications*, 398(2), pp.254–259.

Cheng, X. & Roberts, R.J., 2001. AdoMet-dependent methylation , DNA methyltransferases and base flipping. *Nucleic Acids Research*, 29(18), pp.3784–3795.

Christian, M., Cermak, T., Doyle, E.L., Schmidt, C., Zhang, F., Hummel, A., Bogdanove, A.J., Voytas, D.F., 2010. Targeting DNA Double-Strand Breaks with TAL Effector Nucleases. *Genetics*, 186(2), pp.757–761.

Clark, T.A., Murray, I.A., Morgan, R.D., Kislyuk, A.O., Spittle, K.E., Boitano, M., Fomenkov, A., Roberts, R.J., Korlach, J., 2012. Characterization of DNA methyltransferase specificities using single-molecule , real-time DNA sequencing. *Nucleic acids research*, 40(4), p.e29.

Davies, G.P., Martin, I., Sturrock, S.S., Cronshaw, A., Murray, N.E., Dryden, D.T., 1999. On the structure and operation of type I DNA restriction enzymes. *Journal of Molecular Biology*, 290(2), pp.565–79.

Doudna, J.A. & Charpentier, E., 2014. The new frontier of genome engineering with CRISPR-Cas9. *Science*, 346(6213).

Dudev, T. & Lim, C., 2014. Competition among Metal Ions for Protein Binding Sites: Determinants of Metal Ion Selectivity in Proteins. *Chemical Reviews*, 114(1), pp.538–556.

Dupureur, C.M., 2008. Roles of metal ions in nucleases. *Current Opinion in Chemical Biology*, 12(2), pp.250–255.

Durai, S., Mani, M., Kandavelou, K., Wu, J., Porteus, M.H., Chandrasegaran, S., 2005. Zinc finger nucleases: custom-designed molecular scissors for genome engineering of plant and mammalian cells. *Nucleic Acids Research* , 33(18), pp.5978–5990.

Eid, J., Bjornson, K.P., Chaudhuri, B.P., Cicero, R.L., Flusberg, B.A., Gray, J.J., Holden, D., Saxena, R., Wegener, J., Turner, S.W., 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, 323(January), pp.133–138.

Ellrott, K., Kasarjian, J.K., Jiang, T., Ryu, J., 2002. Restriction enzyme recognition sequence search program. *Biotechniques*, 33(6), pp.1322–1326.

Endlich, B. & Linn, S., 1985. The DNA restriction endonuclease of Escherichia coli B. I. Studies of the DNA translocation and the ATPase activities. *Journal of Biological Chemistry* , 260(9), pp.5720–5728.

Estabrook, R.A., Lipson, R., Hopkins, B., Reich, N., 2004. The Coupling of Tight DNA Binding and Base Flipping. *Journal of Biological Chemistry*, 279(30), pp.31419–31428.

Feil, E.J., Cooper, J.E., Grundmann, H., Robinson, D.A., Enright, M.C., Berendt, T., Peacock, S.J., Smith, J.M., Murphy, M., Spratt, B.G., Moore, C.E., Day, N.P.J., 2003. How Clonal Is Staphylococcus aureus? *Journal of Bacteriology*, 185(11), pp.3307–3316.

Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T.A., Korlach, J., Turner, S.W., 2010. Direct detection of DNA methylation during single-molecule , real-time sequencing. *Nature Publishing Group*, 7(6), pp.461–465.

Fontecave, M., Atta, M. & Mulliez, E., 2004. S-adenosylmethionine: nothing goes to waste. *Trends in Biochemical Sciences*, 29(5), pp.243–249.

Furuta, Y., Namba-Fukuyo, H., Shibata, T.F., Nishiyama, T., Shigenobu, S., Suzuki, Y., Sugano, S., Hasebe, M., Kobayashi, I., 2014. Methylome Diversification through Changes in DNA Methyltransferase Sequence Specificity. *PLoS Genet*, 10(4), p.e1004272.

Furuta, Y. & Kobayashi, I., 2012. Mobility of DNA sequence recognition domains in DNA methyltransferases suggests epigenetics- driven adaptive evolution. *Mobile Genetic Elements*, 2(6), pp.292–296.

Gandon, S., 2016. Why Be Temperate: Lessons from Bacteriophage λ *Trends in Microbiology*, 24(5), pp.356–365.

Gong, W., O'Gara, M., Blumenthal, R.M., Cheng, X., 1997. Structure of PvuII DNA-(cytosine N4) methyltransferase, an example of domain permutation and protein fold assignment. *Nucleic Acids Research*, 25(14), pp.2702–2715.

Gowda, S., Mathew, B.B., Sudhamani, C.N., Bhojya Naik, H.S., 2014. Mechanism of DNA Binding and Cleavage. *Biomedicine and Biotechnology*, 2(1), pp.1–9.

Graille, M., Stura, E.A., Corper, A.L., Sutton, B.J., Taussig, M.J., Charbonnier, J.B., Silverman, G.J., 2000. Crystal structure of a Staphylococcus aureus protein A domain complexed with the Fab fragment of a human IgM antibody: Structural basis for recognition of B-cell receptors and superantigen activity. *Proceedings of the National Academy of Sciences of the United States of America*, 97(10), pp.5399–5404.

Halford, S.E., Catto, L.E., Pernstich, C., Rusling, D.A., Sanders, K.L., 2011. The reaction mechanism of FokI excludes the possibility of targeting zinc finger nucleases to unique DNA sites. *Biochemical Society Transactions*, 39(2), pp.584–588.

Harmsen, D., Claus, H., Witte, W., Rothgänger, J., Claus, H., Turnwald, D., Vogel, U., 2003. Typing of Methicillin-Resistant Staphylococcus aureus in a University Hospital Setting by Using Novel Software for spa Repeat Determination and Database Management. *Journal of Clinical Microbiology*, 41(12), pp.5442–5448.

Heiro, M., Helenius, H., Mäkilä, S., Hohenthal, U., Savunen, T., Engblom, E., Nikoskelainen, J., Kotilainen, P., 2006. Infective endocarditis in a Finnish teaching hospital: a study on 326 episodes treated during 1980–2004. *Heart*, 92(10), pp.1457–1462.

Holubová, I., Vejsadová, Š., Firman, K., Weiserová, M., 2004. Cellular localization of Type I restriction-modification enzymes is family dependent. *Biochemical and biophysical research communications*, 319(2), pp.375–80.

Hong, J.-S., Smith, G.R. & Ames, B.N., 1971. Adenosine 3′:5′-Cyclic Monophosphate Concentration in the Bacterial Host Regulates the Viral Decision between Lysogeny and Lysis. *Proceedings of the National Academy of Sciences of the United States of America*, 68(9), pp.2258–2262.

Horton, J.R., Liebert, K., Bekes, M., Jeltsch, A., Cheng, X., 2006. Structure and substrate recognition of the Escherichia coli DNA adenine methyltransferase. *Journal of Molecular Biology*, 358(2), pp.559–570.

Janscak, P., MacWilliams, M.P., Sandmeier, U., Nagaraja, V., Bickle, T.A., 1999. DNA translocation blockage, a general mechanism of cleavage site selection by type I restriction enzymes. *The EMBO Journal*, 18(9), pp.2638–2647.

Janscak, P. & Bickle, T.A., 1998. The DNA recognition subunit of the type IB restriction-modification enzyme EcoAI tolerates circular permutions of its polypeptide chain1. *Journal of Molecular Biology*, 284(4), pp.937–948.

Jeltsch, A., Christ, F., Fatemi, M., Roth, M., 1999. On the Substrate Specificity of DNA Methyltransferases. *The Journal of Biological Chemistry*, 274(28), pp.19538–19544.

Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices1 . *Journal of Molecular Biology*, 292(2), pp.195–202.

Jurėnaitė-urbanavičienė, A.S., Serksnaite, J., Kriukiene, E., Giedriene, J., Venclovas, C., Lubys, A., 2016. Generation of DNA cleavage specificities of type II restriction endonucleases by reassortment of target recognition domains. *Proceedings of the National Academy of Science of the United States of America*, 104(25), pp.10358–10363.

Kanwar, N., Roberts, G.A., Cooper, L.P., Stephanou, A.S., Dryden, D.T., 2016. The evolutionary pathway from a biologically inactive polypeptide sequence to a folded, active structural mimic of DNA. *Nucleic Acids Research*, 44(9), pp.4289–4303.

Kaur, P., Plochberger, B., Costa, P., Cope, S.M., Vaiana, S.M., Lindsay, S., 2012. Hydrophobicity of methylated DNA as a possible mechanism for gene silencing. *Physical Biology*, 65001(9), pp.1–8.

Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., Sternberg, M.J., 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protocols*, 10(6), pp.845–858.

Kennaway, C.K., Taylor, J.E., Song, C.F., Potrzebowski, W., Nicholson, W., White, J.H., Swiderska, A., Obarska-Kosinska, A., Callow, P., Cooper, L.P., Roberts, G.A., Artero, J., Bujnicki, J.M., Trinick, J., Kneale, G.G., Dryden, D.T.F., 2012. Structure and operation of the DNA-translocating type I DNA restriction enzymes. *Genes & development*, 26(1), pp.92–104.

Kennaway, C.K., Obarska-Kosinska, A., White, J.H., Tuszynska, I., Cooper, L.P., Bujnicki, J.M., Trinick, J., Dryden, D.T.F., 2009. The structure of M.EcoKI Type I DNA methyltransferase with a DNA mimic antirestriction protein. *Nucleic Acids Research*, 37(3), pp.762–770.

Klimašuaskas, S., Nelson, J.L. & Roberts, R.J., 1991. The sequence specificity domain of cytosine-C5 methylases. *Nucleic Acids Research* , 19(22), pp.6183–6190.

Klug, A., 2010. The discovery of zinc fingers and their development for practical applications in gene regulation and genome manipulation. *Quarterly Reviews of Biophysics*, 43(1), pp.1–21.

Kluytmans, J., van Belkum, A. & Verbrugh, H., 1997. Nasal carriage of Staphylococcus aureus: epidemiology, underlying mechanisms, and associated risks. *Clinical Microbiology Reviews*, 10(3), pp.505–520.

Kneale, G.G., 1994. A Symmetrical Model for the Domain Structure of Type I DNA Methyltransferases. *Journal of Molecular Biology*, 243(1), pp.1–5.

Kobayashi, I., 2001. Behavior of restriction–modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Research*, 29(18), pp.3742–3756.

Korlach, J. & Turner, S.W., 2012. Going beyond five bases in DNA sequencing. *Current Opinion in Structural Biology*, 22(3), pp.251–261.

Kostrewa, D. & Winkler, F.K., 1995. Mg^{2+} Binding to the Active Site of EcoRV Endonuclease: A Crystallographic Study of Complexes with Substrate and Product DNA at 2-.ANG. Resolution. *Biochemistry*, 34(2), pp.683–696.

Kovall, R.A. & Matthews, B.W., 1998. Structural, functional, and evolutionary relationships between λ-exonuclease and the type II restriction endonucleases. *Proceedings of the National Academy of Science of the United States of America*, 95(14), pp.7893–7897.

Kovall, R. & Matthews, B.W., 1997. Toroidal Structure of λ-Exonuclease. *Science*, 277(5333), pp.1824–1827.

Kubik, G. & Summerer, D., 2016. TALEored Epigenetics : A DNA-Binding Scaffold for Programmable Epigenome Editing and Analysis. *ChemBioChem*, 10(1002), pp.975–80.

Kulkarni, M., Nirwan, N., vanAeist, K., Szczelkun, M.D., Saikrishnan, K., 2016. Structural insights into DNA sequence recognition by Type ISP restriction-modification enzymes. *Nucleic Acids Research*, 44(9), pp.4396–4408.

Kumar, S., Cheng, X., Klimasauskas, S., Mi, S., Posfai, J., Roberts, R.J., Wilson, G.G., 1994. The DNA (cytosine-5) methyltransferases. *Nucleic Acids Research*, 22(1), pp.1–10.

Kushner, S.R., 1974. Differential Thermolability of Exonuclease and Endonuclease Activities of the recBC Nuclease Isolated from Thermosensitive recB and recC Mutants. *Journal of Bacteriology*, 120(3), pp.1219–1222.

Lapkouski, M., Panjikar, S., Janscak, P., Smatanova, I.K., Carey, J., Ettrich, R., Csefalvay, E., 2009. Structure of the motor subunit of type I restriction-modification complex EcoR124I. *Nature Structural & Molecular Biology*, 16(1), pp.94–5.

Lederberg, E.M. & Lederberg, J., 1953. GENETIC STUDIES OF LYSOGENICITY IN ESCHERICHIA COLI. *Genetics*, 38(1), pp.51–64.

Lekkerkerk, W.S.N., van Wamel, W.J., Snijders, S.V., Willems, R.J., van Duijkeren, E., Broens, E.M., Wagenaar, J.A., Lindsay, J.A., Vos, M.C., 2015. What Is the Origin of Livestock-Associated Methicillin-Resistant Staphylococcus aureus Clonal Complex 398 Isolates from Humans without Livestock Contact? An Epidemiological and Genetic Analysis B. W. Fenwick, ed. *Journal of Clinical Microbiology*, 53(6), pp.1836–1841.

Li, L., Wu, L.P. & Chandrasegaran, S., 1992. Functional domains in Fok I restriction endonuclease. *Proceedings of the National Academy of Science of the United States of America*, 89(May), pp.4275–4279.

Linder, P., Lasko, P.F., Ashburner, M., Leroy, P., Nielsen, P.J., Nishi, K., Schnier, J., Slonimski, P.P., 1989. Birth of the D-E-A-D box. *Nature*, 337(12), pp.121–122.

Lindsay, J.A., 2010. Genomic variation and evolution of Staphylococcus aureus. *International Journal of Medical Microbiology*, 300(2–3), pp.98–103.

Lindsay, J.A., 2013. Hospital-associated MRSA and antibiotic resistance—What have we learned from genomics? *International Journal of Medical Microbiology*, 303(6–7), pp.318–323.

Lindsay, J.A., Moore, C.E., Day, N.P., Peacock, S.J., Witney, A.A., Stabler, R.A., Husain, S.E., Butcher, P.D., Hinds, J., 2006. Microarrays Reveal that Each of the Ten Dominant Lineages of Staphylococcus aureus Has a Unique Combination of Surface-Associated and Regulatory Genes . *Journal of Bacteriology*, 188(2), pp.669–676.

Lindsay, J.A., 2014. Staphylococcus aureus genomics and the impact of horizontal gene transfer. *International Journal of Medical Microbiology*, 304(2), pp.103–109.

Liu, Q. & Wang, J.C., 2016. Similarity in the Catalysis of DNA Breakage and Rejoining by Type IA and IIA DNA Topoisomerases. *Proceedings of the National Academy of Science of the United States of America*, 96(3), pp.881–886.

Loenen, W.A.M. & Raleigh, E.A., 2014. The other face of restriction: modification-dependent enzymes. *Nucleic Acids Research*, 42(1), pp.56–69.

van Loo, I., Huijsdens, X., Tiemersma, E., de Neeling, A., van de Sande-Bruinsma, N., Beaujean, D., Voss, A., Kluytmans, J., 2007. Emergence of Methicillin-Resistant Staphylococcus aureus of Animal Origin in Humans. *Emerging Infectious Diseases*, 13(12), pp.1834–1839.

MacWilliams, M.P. & Bickle, T.A., 1996. Generation of new DNA binding specificity by truncation of the type IC EcoDXXI hsdS gene. *The EMBO Journal*, 15(17), pp.4775–4783.

Madsen, A. & Josephsen, J., 2001. The LlaGI restriction and modification system of Lactococcus lactis W10 consists of only one single polypeptide. *FEMS Microbiology Letters*, 200(1), pp.91–96.

Makovets, S., Powell, L.M., Titheradge, A.J., Blakely, G.W., Murray, N.E., 2004. Is modification sufficient to protect a bacterial chromosome from a resident restriction endonuclease? *Molecular Microbiology*, 51(1), pp.135–147.

Marshall, J.J.T., Gowers, D.M. & Halford, S.E., 2007. Restriction endonucleases that bridge and excise two recognition sites from DNA. *Journal of Molecular Biology*, 367(2), pp.419–31.

Marshall, J.J.T. & Halford, S.E., 2010. The Type IIB restriction endonucleases. *Biochemical Society Transactions*, 38, pp.410–416.

McCarthy, A.J., van Wamel, W., Vandendriessche, S., Larsen, J., Denis, O., Garcia-Graells, C., Uhlemann, A., Lowy, F.D., Skov, R., Lindsay, J.A., 2012. Staphylococcus aureus CC398 Clade Associated with Human-to-Human Transmission. *Applied and Environmental Microbiology*, 78(24), pp.8845–8848.

McCarthy, A.J., Witney, A.A. Gould, K.A., Moodley, A., Guardabassi, L., Voss, A., Denis, O., Broens, E.M., Hinds, J., Lindsay, J.A., 2011. The Distribution of Mobile Genetic Elements (MGEs) in MRSA CC398 Is Associated with Both Host and Country. *Genome Biology and Evolution* , 3, pp.1164–1174.

McClelland, S.E., Dryden, D.T.F. & Szczelkun, M.D., 2005. Continuous Assays for DNA Translocation Using Fluorescent Triplex Dissociation: Application to Type I Restriction Endonucleases. *Journal of Molecular Biology*, 348(4), pp.895–915.

McMahon, S., Roberts, G.A., Johnson, K.A., Cooper, L.P., Liu, H., White, J.H., Carter, L.G., Sanghvi, B., Oke, M., Walkinshaw, M.D., Blakely, G.W., Naismith, J.H., Dryden, D.T.F., 2009. Extensive DNA mimicry by the ArdA anti-restriction protein and its role in the spread of antibiotic resistance. *Nucleic acids research*, 37(15), pp.4887–97.

van der Mee-Marquet, N.L., Corvaglia, A., Haenni, M., Bertrand, X., Franck, J.B., Kluytmans, J., Girard, M., Quentin, R., François, P., 2014. Emergence of a novel subpopulation of CC398 Staphylococcus aureus infecting animals is a serious hazard for humans. *Frontiers in Microbiology*, 5, p.652.

Meng, H., Cao, Y., Qin, J., Song, X., Zhang, Q., Shi, Y., Cao, L., 2015. DNA Methylation, Its Mediators and Genome Integrity. *International Journal of Biological Sciences*, 11(5), pp.604–617.

Mierzejewska, K., Bochtler, M. & Czapinska, H., 2016. On the role of steric clashes in methylation control of restriction endonuclease activity. *Nucleic Acids Research*, 44(1), pp.485–495.

Miller, J., McLachlan, A.D. & Klug, A., 1985. Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *The EMBO Journal*, 4(6), pp.1609–1614.

Mokrishcheva, M.L., Solonin, A.S. & Nikitin, D. V, 2011. Fused eco29kIR- and M genes coding for a fully functional hybrid polypeptide as a model of molecular evolution of restriction-modification systems. *BMC Evolutionary Biology*, 11(35).

Murray, I.A., Clark, T.A., Morgan, R.D., Boitano, M., Anton, B.P., Luong, K., Fomenkov, A., Turner, S.W., Korlach, J., Roberts, R.J., 2012. The methylomes of six bacteria. *Nucleic Acids Research*, 40(22), pp.11450–11462.

Murray, N.E., Daniel, A.S., Cowan, G.M., Sharp, P.M., 1993. Conservation of motifs within the unusually variable polypeptide sequences of type I restriction and modification enzymes. *Molecular Microbiology*, 9(1), pp.133–143.

Murray, N.E., 2000. Type I Restriction Systems : Sophisticated Molecular Machines (a Legacy of Bertani Type I Restriction Systems : Sophisticated Molecular Machines (a Legacy of Bertani and Weigle). *Microbiology and Molecular Biology Reviews*, 64(2), pp.412–34.

Pabo, C.O. & Sauer, R.T., 1984. Protein-DNA Recognition. *Annual Review of Biochemistry*, 53(1), pp.293–321.

Palermo, G., Cavalli, A., Klein, M.L., Alfonso-Prieto, M., Dal Peraro, M., De Vivo, M., 2015. Catalytic Metal Ions and Enzymatic Processing of DNA and RNA. *Accounts of Chemical Research*, 48(2), pp.220–228.

Parrish, J.Z. & Xue, D., 2006. Cuts can kill: the roles of apoptotic nucleases in cell death and animal development. *Chromosoma*, 115(2), pp.89–97.

Pavletich, N.P. & Pabo, C.O., 1991. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science*, 252(5007), pp.809–817.

Piccolo, F.M. & Fisher, A.G., 2014. Getting rid of DNA methylation. *Trends in Cell Biology*, 24(2), pp.136–143.

Pingoud, A., Fuxreiter, M., Pingoud, V., Wende, W., 2005. Type II restriction endonucleases: structure and mechanism. *Cellular and Molecular Life Sciences*, 62(6), pp.685–707.

Pingoud, A. & Jeltsch, A., 2001. Structure and function of type II restriction endonucleases. *Nucleic Acids Research*, 29(18), pp.3705–3727.

Pingoud, A., Wilson, G.G. & Wende, W., 2014. Type II restriction endonucleases— a historical perspective and more. *Nucleic Acids Research*, 42(12), pp.7489–7527.

Poole, A., Penny, D. & Sjöberg, B., 2001. Confounded cytosine! Tinkering and the evolution of DNA. *Nature Reviews Molecular Cell Biology*, 2(February), pp.1–5.

Powell, L.M., Lejeune, E., Hussain, F.S., Cronshaw, A.D., Kelly, S.M., Price, N.C., Dryden, D.T.F., 2003. Assembly of EcoKI DNA methyltransferase requires the C-terminal region of the HsdM modification subunit. *Biophysical chemistry*, 103(2), pp.129–37.

Powell, L.M. & Murray, N.E., 1995. S-adenosyl methionine alters the DNA contacts of the EcoKI methyltransferase. *Nucleic Acids Research*, 23(6), pp.967–974.

Prakash-Cheng, A. & Ryu, J., 1993. Delayed expression of in vivo restriction activity following conjugal transfer of Escherichia coli hsdK (restriction-modification) genes. *Journal of Bacteriology*, 175(15), pp.4905–4906.

Prober, J.M., Trainor, G.L., Dam, R.J., Hobbs, F.W., Robertson, C.W., Zagursky, R.J., Cocuzza, A.J., Jensen, M.A., Baumeister, K., 1987. A System for Rapid DNA Sequencing with Fluorescent Chain-Terminating Dideoxynuclcotides. *Science*, 238(4825), pp.336–341.

Ramchandani, S., Bhattacharya, S.K., Cervoni, N., Szyf, M., 1999. DNA methylation is a reversible biological signal. *Proceedings of the National Academy of Science of the United States of America*, 96(May), pp.6107–6112.

Rao, D.N., Dryden, D.T.F. & Bheemanaik, S., 2014. Type III restriction-modification enzymes: a historical perspective. *Nucleic Acids Research* , 42(1), pp.45–55.

Rasmussen, R.V., Fowler, V.G. Jr., Skov, R., Bruun, N.E., 2011. Future challenges and treatment of Staphylococcus aureus bacteremia with emphasis on MRSA. *Future microbiology*, 6(1), pp.43–56.

Rasool, M., Malik, A., Naseer, M.I., Manan, A., Ansari, S.A., Begum, I., Qazi, M.H., Pushparaj, P.N., Abuzenadah, A.M., Al-Qatani, M.H., Kamal, M.A., Gan, S.H., 2015. The role of epigenetics in personalized medicine: challenges and opportunities. *BMC Medical Genomics*, 8(Suppl 1), pp.1–8.

Ratel, D. & Ravanat, J., 2006. N6-methyladenine: the other methylated base of DNA. *Bioessays*, 28(3), pp.309–315.

Roberts, G.A., Houston, P.J., White, J.H., Chen, K., Stephanou, A.S., Cooper, L.P., Dryden, D.T., Lindsay, J.A., 2013. Impact of target site distribution for Type I restriction enzymes on the evolution of methicillin-resistant Staphylococcus aureus (MRSA) populations. *Nucleic Acids Research*, 41(15), pp.7472–7484.

Roberts, G.A., Chen, K., Bower, E.K.M., Madrzak, J., Woods, A., Barker, A.M., Cooper, L.P., White, J.H., Blakely, G.W., Manfield, I., Dryden, D.T.F., 2013. Mutations of the domain forming the dimeric interface of the ArdA protein affect dimerization and antimodification activity but not antirestriction activity. *The Febs Journal*, 280(19), pp.4903–4914.

Roberts, G.A., Chen, K., Cooper, L.P., White, J.H., Blakely, G.W., Dryden, D.T.F., 2012. Removal of a frameshift between the hsdM and hsdS genes of the EcoKI Type IA DNA restriction and modification system produces a new type of system and links the different families of Type I systems. *Nucleic Acids Research*, 40(21), pp.10916–24.

Roberts, G.A., Cooper, L.P., White, J.H., Su, T., Zipprich, J.T., Geary, P., Kennedy, C., Dryden, D.T.F., 2011. An investigation of the structural requirements for ATP hydrolysis and DNA cleavage by the EcoKI Type I DNA restriction and modification enzyme. *Nucleic Acids Research*, 39(17), pp.7667–76.

Roberts, R.J., Belfort, M., Bestor, T., Bhagwat, A.S., Bickle, T.A., Bitinaite, J., Blumenthal, R.M., Degtyarev, S.K., Dryden, D.T., Dybvig, K., Firman, K., Gromova, E.S., Gumport, R.I., Halford, S.E., Hattman, S., Heitman, J., Hornby, D.P., Janulaitis, A., Jeltsch, A., Josephsen, J., Kiss, A., Klaenhammer, T.R., Kobayashi, I., Kong, H., Krüger, D.H., Lacks, S., Marinus, M.G., Miyahara, M., Morgan, R.D., Murray, N.E., Nagaraja, V., Piekarowicz, A., Pingoud, A., Raleigh, E., Rao, N., Repin, V.E., Selker, E.U., Shaw, P.C., Stein, D.C., Stoddard, B.L., Szybalski, W., Trautner, T.A., Van Etten, J.L., Vitor, J.M., Wilson, G.G., Xu, S.Y., 2003. A nomenclature for restriction enzymes , DNA methyltransferases , homing endonucleases and their genes. *Nucleic Acids Research*, 31(7), pp.1805–1812.

Roberts, R.J., 2005. How restriction enzymes became the workhorses of molecular biology. *Proceedings of the National Academy of Sciences* , 102(17), pp.5905–5908.

Roberts, R.J., Vincze, T., Posfai, J., Macelis, D., 2015. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Research*, 43(Database issue), pp.D298–D299.

Roberts, R.J., Carneiro, M.O. & Schatz, M.C., 2013. The advantages of SMRT sequencing. *Genome Biology*, 14(405), pp.6–9.

Roberts, R.J. & Cheng, X., 1998. BASE FLIPPING. *Annual Review of Biochemistry*, 67, pp.181–98.

Robertson, K.D., 2005. Dna methylation and human disease. *Nature Reviews Genetics*, 6(August), pp.597–610.

Sáez-Llorens, X. & McCracken Jr, G.H., 2003. Bacterial meningitis in children. *The Lancet*, 361(9375), pp.2139–2148.

Salyan, M.E.K., Pedicord, D.L., Bergeron, L., Mintier, G.A., Hunihan, L., Kuit, K., Balanda, L.A., Robertson, B.J., Feder, J.N., Westphal, R., Shipkova, P.A., Blat, Y., 2006. A general liquid chromatography/mass spectroscopy-based assay for detection and quantitation of methyltransferase activity. *Analytical Biochemistry*, 349(1), pp.112–117.

Sambrook, J.F. & Russell, D.W., 2001. *Molecular Cloning, A Laboratory Manual* Third edit., New York: Cold Spring Harbor Laboratory Press.

Sanger, F., Nicklen, S. & Coulson, A.R., 1977. DNA sequencing with chain-terminating. *Proceedings of the National Academy of Science of the United States of America*, 74(12), pp.5463–5467.

Saravanan, M., Vasu, K. & Nagaraja, V., 2016. Evolution of sequence specificity in a restriction endonuclease by a point mutation. *Proceedings of the National Academy of Science of the United States of America*, 105(30), pp.10344–10347.

Scavetta, R.D., Thomas, C.B., Walsh, M.A., Szegedi, S., Joachimiak, A., Gumport., R.I., Churchill, M.E., 2000. Structure of Rsr I methyltransferase, a member of the N6-adenine β class of DNA methyltransferases. *Nucleic Acids Research*, 28(20), pp.3950–3961.

Schneider, C.A., Rasband, W.S. & Eliceiri, K.W., 2012. NIH Image to ImageJ: 25 years of image analysis. *Nat Meth*, 9(7), pp.671–675.

Schulte-Frohlinde, D., 1987. Biological consequences of strand breaks in plasmid and viral DNA. *British Journal of Cancer*, 55(Suppl. VIII), pp.129–133.

Serfiotis-Mitsa, D., Roberts, G.A., Cooper, L.P., White, J.H., Nutley, M., Cooper, A., Blakely, G.W., Dryden, D.T., 2008. The Orf18 Gene Product from Conjugative Transposon Tn916 Is an ArdA Antirestriction Protein that Inhibits Type I DNA Restriction–Modification Systems. *Journal of Molecular Biology*, 383(5), pp.970–981.

Serfiotis-Mitsa, D., Herbert, A.P., Roberts, G.A., Soares, D.C., White, J.H., Blakely, G.W., Uhrin, D., Dryden, D.T., 2010. The structure of the KlcA and ArdB proteins reveals a novel fold and antirestriction activity against Type I DNA restriction systems in vivo but not in vitro. *Nucleic Acids Research*, 38(5), pp.1723–1737.

Shendure, J.A., Porreca, G.J., Church, G.M., Gardner, A.F., Hendrickson, C.L., Kieleczawa, J., Slatko, B.E., 2011. Overview of DNA Sequencing Strategies. *Current protocols in Molecular Biology*, Supplement(October), pp.1–23.

Shieh, F. & Reich, N.O., 2007. AdoMet-dependent Methyl-transfer : Glu 119 Is Essential for DNA C5-Cytosine Methyltransferase M . HhaI. *Journal of Molecular Biology*, 9(37), pp.1157–1168.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWillaims, H., Remmert, M., Söding, J., Thomson, J.D., Higgins, D.G., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7, p.539.

Sigman, D.S. & Chen, C.B., 1990. Chemical Nucleases: New Reagents in Molecular Biology. *Annual Review of Biochemistry*, 59(1), pp.207–236.

Simons, M. & Szczelkun, M.D., 2011. Recycling of protein subunits during DNA translocation and cleavage by Type I restriction-modification enzymes. *Nucleic Acids Research*, 39(17), pp.7656–7666.

Šišáková, E., Stanley, L.K., Weiserová, M., Szczelkun, M.D., 2008. A RecB-family nuclease motif in the Type I restriction endonuclease EcoR124I. *Nucleic Acids Research*, 36(12), pp.3939–3949.

Šišáková, E., Weiserová, M., Dekker, C., Seidel, R., Szczelkun, M.D., 2008. The Interrelationship of Helicase and Nuclease Domains during DNA Translocation by the Molecular Motor EcoR124I. *Journal of Molecular Biology*, 384(5), pp.1273–1286.

Sitaraman, R., 2016. The Role of DNA Restriction-Modification Systems in the Biology of Bacillus anthracis. *Frontiers in Microbiology*, 7, p.11.

Siwek, W., Czapinska, H., Bochtler, M., Bujnicki, J.M., Skowronek, K., 2012. Crystal structure and mechanism of action of the N6-methyladenine-dependent type IIM restriction endonuclease R.DpnI. *Nucleic Acids Research*, 40(15), pp.7563–72.

Smith, H.O. & Nathans, D., 1973. A Suggested nomenclature for bacterial host modification and restriction systems and their enzymes. *Journal of Molecular Biology*, 81(3), pp.419–423.

Smith, R.M., Josephsen, J. & Szczelkun, M.D., 2009. The single polypeptide restriction–modification enzyme LlaGI is a self-contained molecular motor that translocates DNA loops. *Nucleic Acids Research*, 37(21), pp.7219–7230.

Stefani, S., Chung, D.R., Lindsay, J.A., Friedrich, A.W., Kearns, A.M., Westh, H., Mackenzie, F.M., 2012. Meticillin-resistant Staphylococcus aureus (MRSA): global epidemiology and harmonisation of typing methods. *International Journal of Antimicrobial Agents*, 39(4), pp.273–282.

Stegger, M., Lindsay, J.A., Sørum, M., Gould, K.A., Skov, R., 2010. Genetic diversity in CC398 methicillin-resistant Staphylococcus aureus isolates of different geographical origin. *Clinical Microbiology And Infection: The Official Publication Of The European Society Of Clinical Microbiology And Infectious Diseases*, 16(7), pp.1017–1019.

Stegger, M., Lindsay, J.A., Moodley, A., Skov, R., Broens, E.M., Guardabassi, L., 2011. Rapid PCR Detection of Staphylococcus aureus Clonal Complex 398 by Targeting the Restriction-Modification System Carrying sau1-hsdS1. *Journal of Clinical Microbiology*, 49(2), pp.732–734.

Steitz, T.A. & Steitz, J.A., 1993. A general two-metal-ion mechanism for catalytic RNA. *Proceedings of the National Academy of Science of the United States of America*, 90(July), pp.6498–6502.

Stephanou, A.S., Roberts, G.A., Cooper, L.P., Clarke, D.J., Thomson, A.R., Mackay, L., Nutley, M., Cooper, A., Dryden, D.T.F. 2009. Dissection of the DNA Mimicry of the Bacteriophage T7 Ocr Protein using Chemical Modification. *Journal of Molecular Biology*, 391(3), pp.565–576.

Sträter, N., Lipscomb, W.N., Klabunde, T., Krebs, B., 1996. Two-Metal Ion Catalysis in Enzymatic Acyl- and Phosphoryl-Transfer Reactions. *Angewandte Chemie International Edition in English*, 35(18), pp.2024–2055.

Sun, Q., Huang, S., Wang, X., Zhu, Y., Chen, Z., Chen, D., 2015. N6-methyladenine functions as a potential epigenetic mark in eukaryotes. *Bioessays*, 37(11), pp.1155–1162.

Sung, J.M.-L. & Lindsay, J.A., 2007. Staphylococcus aureus Strains That are Hypersusceptible to Resistance Gene Transfer from Enterococci . *Antimicrobial Agents and Chemotherapy*, 51(6), pp.2189–2191.

Swarthout, J.T., Raisinghani, M. & Cui, X., 2011. Zinc Finger Nucleases: A new era for transgenic animals. *Annals of Neurosciences*, 18(1), pp.25–28.

Taylor, I.A., Davis, K.G., Watts, D., Kneale, G.G., 1994. DNA-binding induces a major structural transition in a type I methyltransferase. *The EMBO Journal*, 13(23), pp.5772–5778.

Taylor, J.E., Callow, P., Swiderska, A., Kneale, G.G., 2010. Structural and functional analysis of the engineered type I DNA methyltransferase EcoR124I(NT). *Journal of Molecular Biology*, 398(3), pp.391–9.

Thompson, C.B., 1995. Apoptosis in the Pathogenesis and Treatment of Disease. *Science*, 267(March), pp.1456–1462.

Tock, M.R. & Dryden, D.T.F., 2005. The biology of restriction and anti-restriction. *Current Opinion in Microbiology*, 8(4), pp.466–472.

Tong, S.Y.C., Davis, J.S., Eichenberger, E., Holland, T.L., Fowler, V.G. Jr., 2015. Staphylococcus aureus Infections: Epidemiology, Pathophysiology, Clinical Manifestations, and Management. *Clinical Microbiology Reviews*, 28(3), pp.603–661.

Tsen, H. & Levene, S.D., 2004. Analysis of Chemical and Enzymatic Cleavage Frequencies in Supercoiled DNA. *Journal of Molecular Biology*, 336(5), pp.1087–1102.

Uyen, N.T., Park, S., Choi, J., Lee, H., Nishi, K., Kim, J., 2009. The fragment structure of a putative HsdR subunit of a type I restriction enzyme from Vibrio vulnificus YJ016: implications for DNA restriction and translocation activity. *Nucleic Acids Research*, 37(20), pp.6960–6969.

Vasu, K. & Nagaraja, V., 2013. Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense. *Microbiology and Molecular Biology Reviews*, 77(1), pp.53–72.

Veiga, H. & Pinho, M.G., 2009. Inactivation of the SauI Type I Restriction-Modification System Is Not Sufficient To Generate Staphylococcus aureus Strains Capable of Efficiently Accepting Foreign DNA . *Applied and Environmental Microbiology*, 75(10), pp.3034–3038.

Wah, D.A., Bitinaite, J., Schildkraut, I., Aggarwal, A.K., 1998. Structure of FokI has implications for DNA cleavage. *Proceedings of the National Academy of Sciences* , 95(18), pp.10564–10569.

Waldron, D.E. & Lindsay, J.A., 2006. Sau1: a Novel Lineage-Specific Type I Restriction-Modification System That Blocks Horizontal Gene Transfer into Staphylococcus aureus and between S. aureus Isolates of Different Lineages. *Journal of Bacteriology*, 188(15), pp.5578–5585.

Walkinshaw, M.D., Taylor, P., Sturrock, S.S., Atanasiu, C., Berge, T., Henderdon, R.M., Edwardson, J.M., Dryden, D.T.F., 2002. Structure of Ocr from Bacteriophage T7, a Protein that Mimics B-Form DNA. *Molecular Cell*, 9(1), pp.187–194.

van Wamel, W.J.B., Rooijakkers, S.H., Ruyken, M., van Kessel, K.P., van Strijp, J.A., 2006. The Innate Immune Modulators Staphylococcal Complement Inhibitor and Chemotaxis Inhibitory Protein of Staphylococcus aureus Are Located on β-Hemolysin-Converting Bacteriophages. *Journal of Bacteriology*, 188(4), pp.1310–1315.

Williams, R.J., 2003. Restriction endonuclease. *Molecular Biotechnology*, 23(3), pp.225–243.

Wilson, G.G. & Murray, N.E., 1991. Restriction and Modification Systems. *Annual Review of Genetics*, 25, pp.585–627.

Wolfenden, R. & Snider, M.J., 2001. The Depth of Chemical Time and the Power of Enzymes as Catalysts. *Accounts of Chemical Research*, 34(12), pp.938–945.

Wu, J.C. & Santi, D. V, 1987. Kinetic and Catalytic Mechanism of HhaI Methyltransferase*. *The Journal of Biological Chemistry*, 262(10), pp.4778–4786.

Xu, Y. & Kool, E.T., 1998. Chemical and enzymatic properties of bridging 5′-S-phosphorothioester linkages in DNA. *Nucleic Acids Research*, 26(13), pp.3159–3164.

Yang, W., 2011. Nucleases: diversity of structure, function and mechanism. *Quarterly Reviews of Biophysics*, 44(1), pp.1–93.

Yang, W., Lee, J.Y. & Nowotny, M., 2006. Making and Breaking Nucleic and Substrate Specificity. *Molecular Cell*, 22(1), pp.5–13.

Yuan, R., Hamilton, D.L. & Burckhardt, J., 1980. DNA translocation by the restriction enzyme from E. coli K. *Cell*, 20(1), pp.237–244.

Zangi, R., Arrieta, A. & Cossío, F.P., 2010. Mechanism of DNA Methylation: The Double Role of DNA as a Substrate and as a Cofactor. *Journal of Molecular Biology*, 400(3), pp.632–644.

Computer Programs Used:

FinchTV 1.4.0 (Geospiza, Inc; Seattle, WA, USA; http://www.geospiza.com

Origin (OriginLab, Northampton, MA)

The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.

Appendix A

Plasmid Sequences:

pJFMSEGFP-

CTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGA
CGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCA
TCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAG
TACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGAACGGCATCAA
GGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACT
ACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTG
AGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCT
GGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGCATCATC
ATCATCATCATTAAGAATTCAGCTTGGCTGTTTTGGCGGATGAGAGAAGATTTTCAGCCT
GATACAGATTAAATCAGAACGCAGAAGCGGTCTGATAAAACAGAATTTGCCTGGCGGCA
GTAGCGCGGTGGTCCCACCTGACCCCATGCCGAACTCAGAAGTGAAACGCCGTAGCGCC
GATGGTAGTGTGGGGTCTCCCCATGCGAGAGTAGGGAACTGCCAGGCATCAAATAAAAC
GAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCT
CCTGAGTAGGACAAATCCGCCGGGAGCGGATTTGAACGTTGCGAAGCAACGGCCCGGAG
GGTGGCGGGCAGGACGCCCGCCATAAACTGCCAGGCATCAAATTAAGCAGAAGGCCATC
CTGACGGATGGCCTTTTTGCGTTTCTACAAACTCTTTTGTTTATTTTTCTAAATACATTCAA
ATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTCAATAATATTGAAAAAGGA
AGAGTATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTT
CCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGT
GCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGC
CCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTA
TCCCGTGTTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGAC
TTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGA
ATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAAC
GATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTC
GCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACC
ACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACT
CTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACT
TCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCG
TGGGTCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGT
TATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGA
TAGGTGCCTCACTGATTAAGCATTGGTAACTGTCAGACCAAGTTTACTCATATATACTTTA
GATTGATTTAAAACTTCATTTTTAATTTAAAAGGATCTAGGTGAAGATCCTTTTTGATAAT
CTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAA
AAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAA
AAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTC
CGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAGCCGT
AGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCC
TGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGAC
GATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCC
AGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAG
CGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGA
ACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGT
CGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAG
CCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTT
GCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTG
AGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAG
GAAGCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACAC
CGCATATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATACA
CTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCCCCGACACCCGCCAACACCCGCTG
ACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCT

CCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCTG
CGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATCCGCG
TCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATG
TTAAGGGCGGTTTTTTCCTGTTTGGTCACTGATGCCTCCGTGTAAGGGGGATTTCTGTTCA
TGGGGGTAATGATACCGATGAAACGAGAGAGGATGCTCACGATACGGGTTACTGATGAT
GAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACTGGCGGTATGGATGCGGCG
GGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGATGTAGGTG
TTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAGGGC
GCTGACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATTCATGTTGTT
GCTCAGGTCGCAGACGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGTGAT
TCATTCTGCTAACCAGTAAGGCAACCCCGCCAGCCTAGCCGGGTCCTCAACGACAGGAG
CACGATCATGCGCACCCGTGGCCAGGACCCAACGCTGCCCGAGATGCGCCGCGTGCGGC
TGCTGGAGATGGCGGACGCGATGGATATGTTCTGCCAAGGGTTGGTTTGCGCATTCACAG
TTCTCCGCAAGAATTGATTGGCTCCAATTCTTGGAGTGGTGAATCCGTTAGCGAGGTGCC
GCCGGCTTCCATTCAGGTCGAGGTGGCCCGGCTCCATGCACCGCGACGCAACGCGGGGA
GGCAGACAAGGTATAGGGCGGCGCCTACAATCCATGCCAACCCGTTCCATGTGCTCGCC
GAGGCGGCATAAATCGCCGTGACGATCAGCGGTCCAATGATCGAAGTTAGGCTGGTAAG
AGCCGCGAGCGATCCTTGAAGCTGTCCCTGATGGTCGTCATCTACCTGCCTGGACAGCAT
GGCCTGCAACGCGGGCATCCCGATGCCGCCGGAAGCGAGAAGAATCATAATGGGGAAG
GCCATCCAGCCTCGCGTCGCGAACGCCAGCAAGACGTAGCCCAGCGCGTCGGCCAGCTT
GCAATTCGCGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAA
ACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGT
ATTGGGCGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTT
CACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGC
GAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGT
CGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGC
ATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCA
TTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCG
CTATCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCG
CCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACC
AGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGT
GTCTGGTCAGAGACATCAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGC
AATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAG
AAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCAC
CACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGC
GTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTT
GTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCG
CGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGA
CACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATT
GACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCACCATTCGATGGTGT
CAACGTAAATGCATGCCGCTTCGCCTTCGCGCGCGAATTGCAAGCTGATCCGGGCTTATC
GACTGCACGGTGCACCAATGCTTCTGGCGTCAGGCAGCCATCGGAAGCTGTGGTATGGCT
GTGCAGGTCGTAAATCACTGCATAATTCGTGTCGCTCAAGGCGCACTCCCGTTCTGGATA
ATGTTTTTTGCGCCGACATCATAACGGTTCTGGCAAATATTCTGAAATGAGCTGTTGACA
ATTAATCATCGGCTCGTATAATGTGTGGAATTGTGAGCGGATAACAATTTCACACAGGAA
ACAGAATTAAGCTTGGCTGCAGGTCGACGGATCCAAGAAGGAGATATACAT

pJF-

CATCATCATCATCATCATTAAGAATTCAGCTTGGCTGTTTTGGCGGATGAGAGAAGATTT
TCAGCCTGATACAGATTAAATCAGAACGCAGAAGCGGTCTGATAAAACAGAATTTGCCT
GGCGGCAGTAGCGCGGTGGTCCCACCTGACCCCATGCCGAACTCAGAAGTGAAACGCCG
TAGCGCCGATGGTAGTGTGGGGTCTCCCCATGCGAGAGTAGGGAACTGCCAGGCATCAA
ATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTG
AACGCTCTCCTGAGTAGGACAAATCCGCCGGGAGCGGATTTGAACGTTGCGAAGCAACG
GCCCGGAGGGTGGCGGGCAGGACGCCCGCCATAAACTGCCAGGCATCAAATTAAGCAGA
AGGCCATCCTGACGGATGGCCTTTTTGCGTTTCTACAAACTCTTTTGTTTATTTTTCTAAAT
ACATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTCAATAATATTG

AAAAAGGAAGAGTATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGC
ATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGA
TCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGA
GAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGG
CGCGGTATTATCCCGTGTTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTC
TCAGAATGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGAC
AGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACT
TCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATC
ATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAG
CGTGACACCACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGA
ACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGC
AGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGC
CGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCG
TATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGA
TCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAACTGTCAGACCAAGTTTACTCAT
ATATACTTTAGATTGATTTAAAACTTCATTTTTAATTTAAAAGGATCTAGGTGAAGATCCT
TTTTGATAATCTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGAC
CCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCT
TGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCA
ACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTCTA
GTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCT
CTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTG
GACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTG
CACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGC
TATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGC
AGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTA
TAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGG
GGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGC
TGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTAC
CGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAG
TGAGCGAGGAAGCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTA
TTTCACACCGCATATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCC
AGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCCCCGACACCCGCCAAC
ACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGT
GACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAG
GCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTC
ATCCGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCG
GGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCACTGATGCCTCCGTGTAAGGGGGATTT
CTGTTCATGGGGGTAATGATACCGATGAAACGAGAGAGGATGCTCACGATACGGGTTAC
TGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACTGGCGGTATGGA
TGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGAT
GTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGT
GCAGGGCGCTGACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATTC
ATGTTGTTGCTCAGGTCGCAGACGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTAT
CGGTGATTCATTCTGCTAACCAGTAAGGCAACCCCGCCAGCCTAGCCGGGTCCTCAACGA
CAGGAGCACGATCATGCGCACCCGTGGCCAGGACCCAACGCTGCCCGAGATGCGCCGCG
TGCGGCTGCTGGAGATGGCGGACGCGATGGATATGTTCTGCCAAGGGTTGGTTTGCGCAT
TCACAGTTCTCCGCAAGAATTGATTGGCTCCAATTCTTGGAGTGGTGAATCCGTTAGCGA
GGTGCCGCCGGCTTCCATTCAGGTCGAGGTGGCCCGGCTCCATGCACCGCGACGCAACGC
GGGGAGGCAGACAAGGTATAGGGCGGCGCCTACAATCCATGCCAACCCGTTCCATGTGC
TCGCCGAGGCGGCATAAATCGCCGTGACGATCAGCGGTCCAATGATCGAAGTTAGGCTG
GTAAGAGCCGCGAGCGATCCTTGAAGCTGTCCCTGATGGTCGTCATCTACCTGCCTGGAC
AGCATGGCCTGCAACGCGGGCATCCCGATGCCGCCGGAAGCGAGAAGAATCATAATGGG
GAAGGCCATCCAGCCTCGCGTCGCGAACGCCAGCAAGACGTAGCCCAGCGCGTCGGCCA
GCTTGCAATTCGCGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGG
GAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTG
CGTATTGGGCGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCC
CTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAG

GCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATC
GTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGC
GCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCT
CATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTC
CGCTATCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACG
CGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGA
CCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGG
GTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACA
GCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCG
AGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACC
ACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGC
GCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAG
TTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCC
CGCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGA
GACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAA
TTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCACCATTCGATGGT
GTCAACGTAAATGCATGCCGCTTCGCCTTCGCGCGCGAATTGCAAGCTGATCCGGGCTTA
TCGACTGCACGGTGCACCAATGCTTCTGGCGTCAGGCAGCCATCGGAAGCTGTGGTATGG
CTGTGCAGGTCGTAAATCACTGCATAATTCGTGTCGCTCAAGGCGCACTCCCGTTCTGGA
TAATGTTTTTTGCGCCGACATCATAACGGTTCTGGCAAATATTCTGAAATGAGCTGTTGA
CAATTAATCATCGGCTCGTATAATGTGTGGAATTGTGAGCGGATAACAATTTCACACAGG
AAACAGAATTAAGCTTGGCTGCAGGTCGACGGATCCAAGAAGGAGATATACAT

pRSFDuet-1-

GGGGAATTGTGAGCGGATAACAATTCCCCTGTAGAAATAATTTTGTTTAACTTTAATAAG
GAGATATACCATGGGCAGCAGCCATCACCATCATCACCACAGCCAGGATCCGAATTCGA
GCTCGGCGCGCCTGCAGGTCGACAAGCTTGCGGCCGCATAATGCTTAAGTCGAACAGAA
AGTAATCGTATTGTACACGGCCGCATAATCGAAATTAATACGACTCACTATAGGGGAATT
GTGAGCGGATAACAATTCCCCATCTTAGTATATTAGTTAAGTATAAGAAGGAGATATACA
TATGGCAGATCTCAATTGGATATCGGCCGGCCACGCGATCGCTGACGTCGGTACCCTCGA
GTCTGGTAAAGAAACCGCTGCTGCGAAATTTGAACGCCAGCACATGGACTCGTCTACTAG
CGCAGCTTAATTAACCTAGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGG
GGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAACCTCAGGCATTTGAGAAGCACAC
GGTCACACTGCTTCCGGTAGTCAATAAACCGGTAAACCAGCAATAGACATAAGCGGCTA
TTTAACGACCCTGCCCTGAACCGACGACAAGCTGACGACCGGGTCTCCGCAAGTGGCACT
TTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGT
ATCCGCTCATGAATTAATTCTTAGAAAAACTCATCGAGCATCAAATGAAACTGCAATTTA
TTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTGTAATGAAGGAGAA
AACTCACCGAGGCAGTTCCATAGGATGGCAAGATCCTGGTATCGGTCTGCGATTCCGACT
CGTCCAACATCAATACAACCTATTAATTTCCCCTCGTCAAAAATAAGGTTATCAAGTGAG
AAATCACCATGAGTGACGACTGAATCCGGTGAGAATGGCAAAAGTTTATGCATTTCTTTC
CAGACTTGTTCAACAGGCCAGCCATTACGCTCGTCATCAAAATCACTCGCATCAACCAAA
CCGTTATTCATTCGTGATTGCGCCTGAGCGAGACGAAATACGCGGTCGCTGTTAAAAGGA
CAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACACTGCCAGCGCATCAACAAT
ATTTTCACCTGAATCAGGATATTCTTCTAATACCTGGAATGCTGTTTTCCCGGGGATCGCA
GTGGTGAGTAACCATGCATCATCAGGAGTACGGATAAAATGCTTGATGGTCGGAAGAGG
CATAAATTCCGTCAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCT
ACCTTTGCCATGTTTCAGAAACAACTCTGGCGCATCGGGCTTCCCATACAATCGATAGAT
TGTCGCACCTGATTGCCCGACATTATCGCGAGCCCATTTATACCCATATAAATCAGCATC
CATGTTGGAATTTAATCGCGGCCTAGAGCAAGACGTTTCCCGTTGAATATGGCTCATACT
CTTCCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATA
TTTGAATGTATTTAGAAAAATAAACAAATAGGCATGCAGCGCTCTTCCGCTTCCTCGCTC
ACTGACTCGCTACGCTCGGTCGTTCGACTGCGGCGAGCGGTGTCAGCTCACTCAAAAGCG
GTAATACGGTTATCCACAGAATCAGGGGATAAAGCCGGAAAGAACATGTGAGCAAAAA
GCAAAGCACCGGAAGAAGCCAACGCCGCAGGCGTTTTTCCATAGGCTCCGCCCCCCTGA
CGAGCATCACAAAAATCGACGCTCAAGCCAGAGGTGGCGAAACCCGACAGGACTATAAA
GATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCT

TACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGC
TGTTGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCC
CCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTA
AGACACGACTTATCGCCACTGGCAGCAGCCATTGGTAACTGATTTAGAGGACTTTGTCTT
GAAGTTATGCACCTGTTAAGGCTAAACTGAAAGAACAGATTTTGGTGAGTGCGGTCCTCC
AACCCACTTACCTTGGTTCAAAGAGTTGGTAGCTCAGCGAACCTTGAGAAAACCACCGTT
GGTAGCGGTGGTTTTTCTTTATTTATGAGATGATGAATCAATCGGTCTATCAAGTCAACG
AACAGCTATTCCGTTACTCTAGATTTCAGTGCAATTTATCTCTTCAAATGTAGCACCTGAA
GTCAGCCCCATACGATATAAGTTGTAATTCTCATGTTAGTCATGCCCCGCGCCCACCGGA
AGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGAGATCCCGGTGCCTAATGA
GTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTG
TCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGG
GCGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCG
CCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAA
TCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTAT
CCCACTACCGAGATGTCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGC
GCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCATTCAG
CATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATC
GGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGA
GACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGAT
GCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCT
GGTCAGAGACATCAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATG
GCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAGAAG
ATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCAC
GCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTG
CAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTT
GTGCCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGT
TTTCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACAC
CGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACT
CTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCGATGGTGTCCGG
GATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAATTAATACGACTCACTATA

The protocol for PCR purification can be found via the following link:

https://www.qiagen.com/gb/resources/resourcedetail?id=390a728a-e6fc-43f7-bf59-b12091cc4380&lang=en

The protocol for plasmid preparation, using the QIAprep Spin Miniprep kit can be found via the following link:

https://www.qiagen.com/gb/resources/resourcedetail?id=89bfa021-7310-4c0f-90e0-6a9c84f66cee&lang=en

Product information and protocols for the Wizard ® Genomic DNA Purification Kit (*Promega*) can be found at:

https://www.promega.co.uk/resources/protocols/technical-manuals/0/wizard-genomic-dna-purification-kit-protocol/

All protocols and publications on the subject of SMRT sequencing by *Pacific Biosciences* can be found on the website, using the following link:

http://www.pacb.com/support/training/

Information and the protocol for using the PD-10 column can be found via the following link:

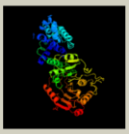
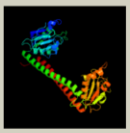
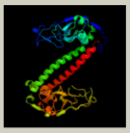
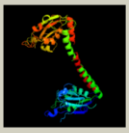
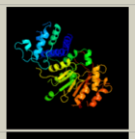
https://www.gelifesciences.com/gehcls_images/GELS/Related%20Content/Files/1314723116657/litdoc52130800BB_20110830191706.pdf

The free to use, gel analysis software, Image J, can be downloaded from the following website:

https://imagej.nih.gov/ij/

Appendix B

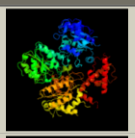
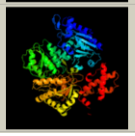
Phyre[2] software aligned the target CC398-1 MTase amino acid sequence with 120 similar protein sequences. Eight of these were chosen, to which sections of target sequence were modelled. Modelling results of these are listed below.

#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	c3lkdB_	Alignment		100.0	49	**PDB header:**transferase **Chain:** B: **PDB Molecule:**type i restriction-modification system **PDBTitle:** crystal structure of the type i restriction-modification2 system methyltransferase subunit from streptococcus3 thermophilus, northeast structural genomics consortium4 target sur80 Run Investigator
8	c1yf2A_	Alignment		100.0	20	**PDB header:**hydrolase regulator **Chain:** A: **PDB Molecule:**type i restriction-modification enzyme, s subunit; **PDBTitle:** three-dimensional structure of dna sequence specificity (s) subunit of2 a type i restriction-modification enzyme and its functional3 implications Run Investigator
9	c1ydxA_	Alignment		100.0	16	**PDB header:**dna binding protein **Chain:** A: **PDB Molecule:**type i restriction enzyme specificity protein mg438; **PDBTitle:** crystal structure of type-i restriction-modification system s subunit2 from m. genitalium Run Investigator
10	c3okgB_	Alignment		100.0	15	**PDB header:**dna binding protein **Chain:** B: **PDB Molecule:**restriction endonuclease s subunits; **PDBTitle:** crystal structure of hsds subunit from thermoanaerobacter2 tengcongensis Run Investigator
11	c2y7cA_	Alignment		100.0	15	**PDB header:**transferase **Chain:** A: **PDB Molecule:**type-1 restriction enzyme ecoki specificity protein; **PDBTitle:** atomic model of the ocr-bound methylase complex from the2 type i restriction-modification enzyme ecoki (m2s1). based3 on fitting into em map 1534. Run Investigator
12	c5hr4J_	Alignment		100.0	19	**PDB header:**hydrolase/dna **Chain:** J: **PDB Molecule:**mmei; **PDBTitle:** structure of type iii restriction-modification enzyme mmei in complex2 with dna has implications for engineering of new specificities Run Investigator
13	c1aqiB_	Alignment		100.0	17	**PDB header:**methyltransferase **Chain:** B: **PDB Molecule:**adenine-n6-dna-methyltransferase taqi; **PDBTitle:** structure of adenine-n6-dna-methyltransferase taqi Run Investigator
14	c1g38A_	Alignment	 Click to download PDB format model	100.0	14	**PDB header:**transferase/dna **Chain:** A: **PDB Molecule:**modification methylase taqi; **PDBTitle:** adenine-specific methyltransferase m. taq i/dna complex Run Investigator

PDB codes for these structures are:

1. 3LKD
8. 1YF2
9. 1YDX
10. 3OKG
11. 2Y7C
12. 5HR4
13. 1AQI
14. 1G38

Phyre[2] software aligned the target CC5 HsdR amino acid sequence with 120 similar protein sequences. Two of these were chosen, to which sections of target sequence were modelled. Modelling results of these are listed below.

#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	c2w00B	Alignment		100.0	39	**PDB header:**hydrolase **Chain:** B: **PDB Molecule:**hsdr; **PDBTitle:** crystal structure of the hsdr subunit of the ecor124i2 restriction enzyme in complex with atp Run Investigator
2	c2w74B	Alignment		100.0	40	**PDB header:**hydrolase **Chain:** B: **PDB Molecule:**type i restriction enzyme ecor124ii r protein; **PDBTitle:** mutant (k220r) of the hsdr subunit of the ecor124i2 restriction enzyme in complex with atp Run Investigator

PDB codes for these structures are:

1. 2W00
2. 4BE7

Appendix C

Plasmid Maps:

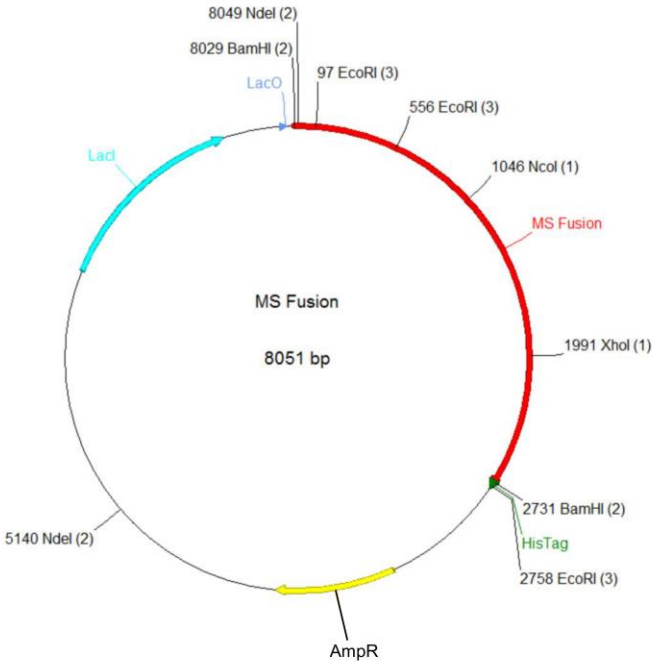


Figure 115: Plasmid map of the MS fusion gene in vector pJF.

Figure 116: Plasmid map of the *hsdM* and HalfSHis genes in vector pJF.

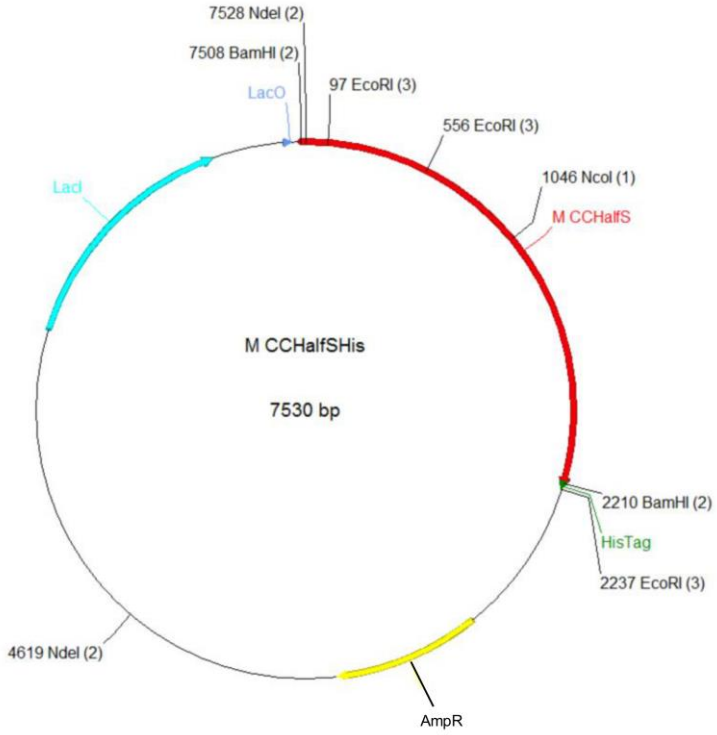


Figure 117: Plasmid map of the *hsdM* and CCHalfSHis genes in vector pJF.

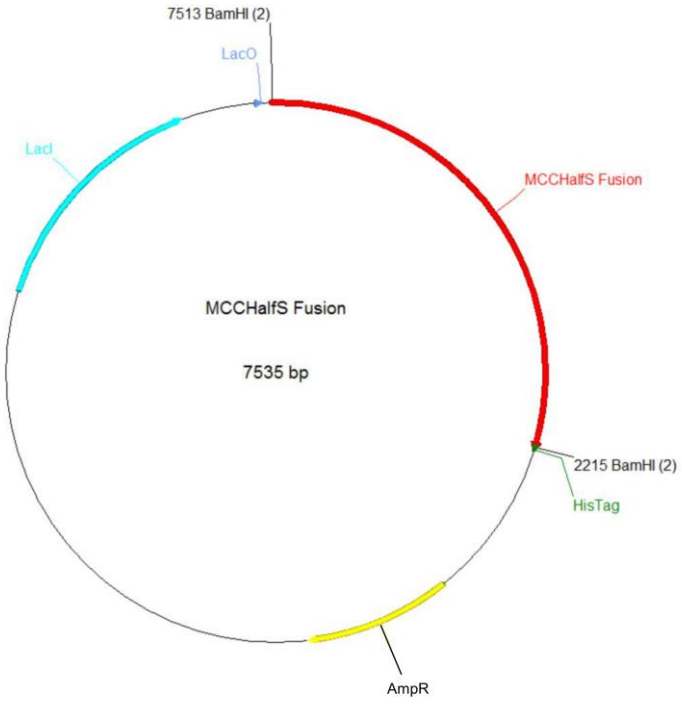


Figure 118: Plasmid map of the MCCHalfS fusion gene in vector pJF.

Appendix D

Image J calculations:

CC398-1

M- 18406.7 ÷ 59454 = 0.310
11663.9 ÷ 59454 = 0.196
16016.6 ÷ 59454 = 0.269
17126.0 ÷ 59454 = 0.288

$\frac{0.310 + 0.196 + 0.269 + 0.288}{4} = 0.266$

$(0.31 - 0.266)^2 = 0.0019$
$(0.196 - 0.266)^2 = 0.0049$
$(0.269 - 0.266)^2 = 0.000009$
$(0.288 - 0.266)^2 = 0.00048$

$\frac{0.0019 + 0.0049 + 0.000009 + 0.0048}{4} = 0.0018$

$\sqrt{0.0018} = 0.04$

S- 5821.5 ÷ 47229 = 0.123
2523.0 ÷ 47229= 0.053
6873.2 ÷ 47229= 0.146
8730.1 ÷ 47229= 0.185

$\frac{0.123 + 0.053 + 0.146 + 0.185}{4} = 0.127$

$(0.123 - 0.127)^2 = 0.000016$
$(0.053 - 0.127)^2 = 0.0055$
$(0.146 - 0.127)^2 = 0.00036$
$(0.185 - 0.127)^2 = 0.0034$

$\frac{0.000016 + 0.0055 + 0.00036 + 0.0034}{4} = 0.0023$

$\sqrt{0.0023} = 0.05$

Ratio = 0.266:0.127

1 ÷ 0.127 = 7.87
0.266 X 7.87 = 2.09
0.04 X 7.87 = 0.31
0.05 X 7.87 = 0.39

Subunit	Intensity	Normalised Intensity	Average	Standard deviation	Rationalised value	Rationalised standard deviation
M	18406.7	0.310	0.2660	0.04	2.09	0.31
	11663.9	0.196				
	16016.6	0.269				
	17126.0	0.288				
S	5821.5	0.123	0.1270	0.05	1.00	0.39
	2523.0	0.053				
	6873.2	0.146				
	8730.1	0.185				

HalfSHis

M- 11110.8 ÷ 59454 = 0.187
11393.1 ÷ 59454 = 0.192
28793.3 ÷ 59454 = 0.484
11908.6 ÷ 59454 = 0.200

$\frac{0.187 + 0.192 + 0.484 + 0.200}{4} = 0.266$

$(0.187 - 0.266)^2 = 0.0062$
$(0.192 - 0.266)^2 = 0.0055$
$(0.484 - 0.266)^2 = 0.0475$
$(0.254 - 0.266)^2 = 0.0001$

$\frac{0.0062 + 0.0055 + 0.0475 + 0.0001}{4} = 0.0148$

$\sqrt{0.0148} = 0.122$

HalfS- 3205.4 ÷ 24930 = 0.129
3158.7 ÷ 24930 = 0.127
15484.1 ÷ 24930 = 0.621
5765.6 ÷ 24930 = 0.231

$\frac{0.129 + 0.127 + 0.621 + 0.231}{4} = 0.277$

$(0.129 - 0.277)^2 = 0.0219$
$(0.127 - 0.277)^2 = 0.0036$
$(0.621 - 0.277)^2 = 0.1183$
$(0.231 - 0.277)^2 = 0.0021$

$\frac{0.\ 0219 + 0.\ 0036 + 0.\ 1183 + 0.\ 0021}{4} = 0.0365$

$\sqrt{0.0365} = 0.191$

Ratio = 0.266:0.277

1 ÷ 0.266 = 3.76
0.277 X 3.75 = 1.04
0.122 X 3.75 = 0.46
0.191 X 3.75 = 0.72

Subunit	Intensity	Normalised Intensity	Average	Standard deviation	Rationalised value	Rationalised standard deviation
M	11110.8	0.187	0.266	0.122	1.00	0.46
	11393.1	0.192				
	28793.3	0.484				
	11908.6	0.200				
HalfS	3205.4	0.129	0.277	0.191	1.04	0.72
	3158.7	0.127				
	15484.1	0.621				
	5765.6	0.231				

CCHalfSHis

M- 31641.0 ÷ 59454 = 0.532
9673.5 ÷ 59454 = 0.163
19027.4 ÷ 59454 = 0.320
15110.0 ÷ 59454 = 0.254

$\frac{0.532 + 0.163 + 0.320 + 0.254}{4} = 0.317$

$(0.532 - 0.317)^2 = 0.0462$
$(0.163 - 0.317)^2 = 0.0237$
$(0.320 - 0.317)^2 = 0.000009$
$(0.254 - 0.317)^2 = 0.0040$

$\frac{0.0462 + 0.0237 + 0.000009 + 0.0040}{4} = 0.0185$

$\sqrt{0.0185} = 0.136$

CCHalfS- 10406.3 ÷ 26020 = 0.400
1047.5 ÷ 26020 = 0.040
4109.5 ÷ 26020 = 0.158
2408.9 ÷ 26020 = 0.093

$\frac{0.400 + 0.040 + 0.158 + 0.093}{4} = 0.173$

$(0.400 - 0.173)^2 = 0.0515$
$(0.040 - 0.173)^2 = 0.0177$
$(0.158 - 0.173)^2 = 0.0002$
$(0.093 - 0.173)^2 = 0.0064$

$\frac{0.0515 + 0.0177 + 0.0002 + 0.0064}{4} = 0.0190$

$\sqrt{0.0190} = 0.138$

Ratio = 0.317:0.173

1 ÷ 0.173 = 5.78
0.317 X 5.78 = 1.83
0.136 X 5.78 = 0.79
0.138 X 5.78 = 0.80

Subunit	Intensity	Normalised Intensity	Average	Standard deviation	Rationalised value	Rationalised standard deviation
M	31641.0	0.532	0.317	0.136	1.83	0.79
	9673.5	0.163				
	19027.4	0.320				
	15110.0	0.254				
CCHalfS	10406.3	0.400	0.173	0.138	1.00	0.80
	1047.5	0.040				
	4109.5	0.158				
	2408.9	0.093				

Appendix E

E.O.P. Calculations:

CC398-1 MTase

Unmodifiedλ Calculations:

1 mL ÷ 100 µL = 10
1516 plaques X 10 X dilution = 15160 X 10^{-5}

= 1516 X 10^6 undiluted phage

Control (None) average- ([1516 + 2630 + 2380 + 2190 +2080] X 10^6) ÷ 5 = 2159.2 X 10^6

$(1516 – 2159.2)^2$ = 413706
$(2630 – 2159.2)^2$ = 221653
$(2380 – 2159.2)^2$ = 48753
$(2190 – 2159.2)^2$ = 949
$(2080 – 2159.2)^2$ = 6273

$\frac{413706 + 221653 + 48753 + 949 + 6273}{5}$ = 138266.8

√138266.8 = 371.8 X 10^6

1 mL ÷ 40 µL = 25
910 plaques X 25 X dilution = 22750 X 10^{-3}

= 22.8 X 10^6 undiluted phage

CC398 MTase + R average- ([22.8 + 80.5 + 379.6 + 512 + 628] X 10^6) ÷ 5 = 324.6 X 10^6

$(22.8 – 324.6)^2$ = 91083
$(80.5 – 324.6)^2$ = 59585
$(379.6 – 324.6)^2$ = 3025
$(512 – 324.6)^2$ = 35119
$(628 – 324.6)^2$ = 92052

$\frac{91083 + 59585 + 3025 + 35119 + 92052}{5}$ = 56172.8

√56172.8 = 237.0 X 10^6

Percentage error

$\frac{371.8}{2159}$ X 100 = 17%

$\frac{237.0}{324.6}$ X 100 = 73%

Combined error

$17^2 + 73^2$ = combined $error^2$ = 941

√5618 = 75.0 %

$\frac{0.15}{100}$ X 75 = 0.11

E.O.P. of Unmodifiedλ

324.6 ÷ 2159.2 = 0.15 (±0.11)

CC398λ Calculations:

CC398 MTase average- ([1600 + 1480 + 2700 + 1800] X 10^6) ÷ 4 = 1895 X 10^6

$(1600 – 1895)^2 = 87025$
$(1480 – 1895)^2 = 172225$
$(2700 – 1895)^2 = 648025$
$(1800 – 1895)^2 = 9025$

$\frac{87025 + 172225 + 648025 + 9025}{4} = 229075$

$\sqrt{229075} = 478.6 \text{ X } 10^6$

CC398 MTase + R average- ([1580 + 1400 + 1780 + 1920] X 10^6) ÷ 4 = 1670 X 10^6

$(1580 – 1670)^2 = 8100$
$(1400 – 1670)^2 = 72900$
$(1780 – 1670)^2 = 12100$
$(1920 – 1670)^2 = 62500$

$\frac{8100 + 72900 + 12100 + 62500}{4} = 38900$

$\sqrt{38900} = 197.2$

Percentage error

$\frac{478.6}{1670} \text{ X } 100 = 29\ \%$ $\frac{197.2}{1895} \text{ X } 100 = 10\ \%$

Combined error

$29^2 + 10^2 = \text{combined error}^2 = 941$

$\sqrt{941} = 30.6\ \%$

$\frac{0.88}{100} \text{ X } 30.6 = 0.27$

E.O.P. of CC398λ

1670 ÷ 1895 = 0.88 (±0.27)

MS fusion MTase:

Unmodifiedλ Calculations:

1 mL / 22.5 µL = 44.4

384 X 44.4 X dilution = 17066.6 X 10^{-5}
= 1706.7 X 10^{-6}

Control (None) average - ([1706.7 + 1920.0 + 2740.0 + 1516.0 + 2630.0] X 10^6)/ 5 = 2102.5 X 10^6

$(1706.7 – 2102.5)^2$ = 156657.6
$(1920 – 2102.5)^2$ = 33306.3
$(2740 – 2102.5)^2$ = 406406.3
$(1516 – 2102.5)^2$ = 343982.3
$(2630 – 2102.5)^2$ = 278256.3

$\frac{156657.6 + 33306.3 + 406406.3 + 343982.3 + 278256.3}{5}$ = 243721.8

$\sqrt{243721.8}$ = 493.7

MSF MTase + R average- ([185.7 + 362.3 + 85 + 150 + 165] X 10^6)/ 5 = 189.6 X 10^6

$(185.7 – 189.6)^2$ = 15.2
$(362.3 – 189.6)^2$ = 29825.3
$(85 – 189.6)^2$ = 10941.2
$(150 – 189.6)^2$ = 1568.2
$(165 – 189.6)^2$ = 605.2

$\frac{15.2 + 29825.3 + 10941.2 + 1568.2 + 605.2}{5}$ = 8591.0

$\sqrt{8591.0}$ = 92.7

Percentage error

$\frac{493.7}{2102.5}$ X 100 = 23.5 % $\frac{92.7}{198.6}$ X 100 = 46.7 %

Combined error

$23.5^2 + 46.7^2$ = combined $error^2$ = 2733.2

$\sqrt{2733.2}$ = 52.3 %

$\frac{0.09}{100}$ X 52.3 = 0.05

E.O.P. of Unmodifiedλ

189.6/ 2102.5 = 0.09 (±0.05)

MSFλ Calculations:

1 mL / 32.3 µL = 30.96

401 plaques X 30.96 X dilution = 12952.3 X 10^{-5}

= 1295.23 X 10^6 undiluted phage

MSF MTase average- ([1295.2 + 1311.4 + 1840 + 1705 + 1460] X 10^6)/ 5 = 1522.3 X 10^6

$(1295.2 – 1522.3)^2$ = 51574.4
$(1311.4 – 1522.3)^2$ = 44478.8
$(1840 – 1522.3)^2$ = 100933.3
$(1705 – 1522.3)^2$ = 33379.3
$(1460 – 1522.3)^2$ = 3881.3

$\frac{51574.4 + 44478.8 + 100933.3 + 33379.3 + 3881.3}{5}$ = 46849.4

√46849.4 = 216.4

MSF MTase + R average- ([1327.3 + 957.5 + 1205 + 1185 + 1235] X 10^6)/ 5 = 1182.0 X 10^6

$(1327.3 – 1182.0)^2$ = 21112.1
$(957.5 – 1182.0)^2$ = 50400.3
$(1205 – 1182.0)^2$ = 529
$(1185 – 1182.0)^2$ = 9
$(1235 – 1182.0)^2$ = 2809

$\frac{21112.1 + 50400.3 + 529 + 9 + 2809}{5}$ = 14971.9

√14971.9 = 122.4

Percentage error

$\frac{216.4}{1522.3}$ X 100 = 14.2 % $\frac{122.4}{1182.0}$ X 100 = 10.4 %

Combined error

$14.2^2 + 10.4^2$ = combined $error^2$ = 309.8

√309.8 = 17.6 %

$\frac{0.78}{100}$ X 17.6 = 0.14

E.O.P. of MSFλ

1182.0/ 1522.32 = 0.78 (±0.14)

HalfSHis MTase:

Unmodifiedλ Calculations:

1 mL / 22.5 µL = 44.4

384 X 44.4 X dilution = 17066.6 X 10^{-5}
= 1706.7 X 10^{-6}

Control (None) average- ([1706.7 + 1920 + 2740 + 1516 + 2630] X 10^6)/ 5 = 2102.5 X 10^6

$(1706.7 – 2102.5)^2 = 156657.6$
$(1920 – 2102.5)^2 = 3306.3$
$(2740 – 2102.5)^2 = 406406.3$
$(1516 – 2102.5)^2 = 343982.3$
$(2630 – 2102.5)^2 = 278256.3$

$\frac{156657.6 + 3306.3 + 406406.3 + 343982.3 + 278256.3}{5} = 237721.8$

$\sqrt{237721.8} = 487.6$

HalfSHis MTase + R average- $([394.6 + 444.1 + 1280 + 900 + 1232.5] \times 10^6)/ 5 = 850.2 \times 10^6$

$(394.6 – 850.2)^2 = 207571.4$
$(444.1 – 850.2)^2 = 164917.2$
$(1280 – 850.2)^2 = 184728.0$
$(900 – 850.2)^2 = 2480.0$
$(1232.5 – 850.2)^2 = 146153.3$

$\frac{207571.4 + 164917.2 + 184728.0 + 2480.0 + 146153.3}{5} = 141170.0$

$\sqrt{141170.0} = 375.7$

Percentage error

$\frac{487.6}{2102.5} \times 100 = 23.2\ \%$ $\frac{375.7}{850.2} \times 100 = 44.2\ \%$

Combined error

$23.2^2 + 44.2^2 = \text{combined error}^2 = 2491.8$

$\sqrt{2491.8} = 49.9\ \%$

$\frac{0.40}{100} \times 49.9 = 0.20$

E.O.P. of Unmodifiedλ

$850.2 / 2102.5 = 0.40 \pm 0.20$

HalfSλ Calculations:

1 mL / 28.5 µL = 35.1

264 plaques X 35.1 X dilution = 9263.1578×10^{-5}

$= 926.3 \times 10^6$ undiluted phage

HalfSHis MTase average- ([926.3 + 1107.8 + 1555.0 + 1620.0 + 1715.0] X 10^6)/ 5 = 1384.8 X 10^6

$(926.3 – 1384.8)^2$ = 210222.3
$(1107.8 – 1384.8)^2$ = 76729.0
$(1555 – 1384.8)^2$ = 28968.0
$(1620 – 1384.8)^2$ = 55319.0
$(1715 – 1384.8)^2$ = 109032.0

$\frac{210222.3 + 76729.0 + 28968.0 + 55319.0 + 109032.0}{5} = 96054.1$

$\sqrt{96054.1} = 309.9$

HalfSHis MTase + R average- ([764.7 + 757.5 + 1075.0 + 1145.0 + 1105.0] X 10^6)/ 5 = 969.4 X 10^6

$(764.7 – 969.4)^2$ = 41902.1
$(757.5 – 969.4)^2$ = 44901.6
$(1075 – 969.4)^2$ = 11151.4
$(1145 – 969.4)^2$ = 30835.4
$(1105 – 969.4)^2$ = 18387.4

$\frac{41902.1 + 44901.6 + 11151.4 + 30835.4 + 18387.4}{5} = 29435.6$

$\sqrt{29435.6} = 171.6$

Percentage error

$\frac{309.9}{1384.8} \times 100 = 22.4\ \%$ $\frac{171.6}{969.4} \times 100 = 17.7\ \%$

Combined error

$22.4^2 + 17.7^2$ = combined $error^2$ = 815.1

$\sqrt{815.1} = 28.5\ \%$

$\frac{0.70}{100} \times 28.5 = 0.20$

E.O.P. of HalfSλ

969.4/ 1384.8 = 0.70 ±0.20

CCHalfSHis MTase:

Unmodifiedλ Calculations:

1 mL / 22.5 µL = 44.4

384 X 44.4 X dilution = 17066.6 X 10^{-5} = 17066.6 X 10^{5}
= 1706.7 X 10^{6} undiluted phage

Control (None) average- ([1706.7 + 1920 + 2740 + 1516 + 2630] X 10^{6})/ 5 = 2102.5 X 10^{6}

$(1706.7 – 2102.5)^2$ = 156657.6
$(1920 – 2102.5)^2$ = 3306.3
$(2740 – 2102.5)^2$ = 406406.3
$(1516 – 2102.5)^2$ = 343982.3
$(2630 – 2102.5)^2$ = 278256.3

$\frac{156657.6 + 3306.3 + 406406.3 + 343982.3 + 278256.3}{5}$ = 237721.8

√237721.8= 487.6

CCHalfSHis MTase + R average- ([727.5 + 852.3 + 1437.5 + 1550.0 + 1490.0] X 10^{6})/ 5 = 1211.5 X 10^{6}

$(727.5 – 1211.5)^2$ = 234256
$(852.3 – 1211.5)^2$ = 129024.6
$(1437.5 – 1211.5)^2$ = 51076
$(1550.0 – 1211.5)^2$ = 114582.3
$(1490.0 – 1211.5)^2$ = 77562.3

$\frac{234256 + 129024.6 + 51076 + 114582.3 + 77562.3}{5}$ = 121300.2

√121300.2 = 348.3

Percentage error

$\frac{487.6}{2102.5}$ X 100 = 23.2 % $\frac{348.3}{1211.5}$ X 100 = 28.7 %

Combined error

$23.2^2 + 28.7^2$ = combined $error^2$ = 1361.9

√1361.9 = 36.9 %

$\frac{0.58}{100}$ X 36.9 = 0.21

E.O.P. of Unmodifiedλ

1211.46/ 2102.54 = 0.58 ±0.21

CCHalfSλ Calculations:

1 mL / 96 µL = 10.42

294 plaques X 10.42 X dilution = 3063.5 X 10^{-6}
= 3063.5 X 10^{6} undiluted phage

CCHalfSHis MTase average- ([3063.5 + 3600.0 + 3710.0 + 4340.0) X 10^{6})/ 4 =
3678.4 X 10^{6}

$(3063.5 – 3678.4)^2$ = 378102.0
$(3600.0 – 3678.4)^2$ = 6146.6
$(3710.0 – 3678.4)^2$ = 998.6
$(4340.0 – 3678.4)^2$ = 437714.6

$\frac{378102.0 + 6146.6 + 998.6 + 437714.6}{4}$ = 205740.5

√205740.5 = 453.6

CCHalfSHis MTase + R average- ([2153.8 + 2740.0 + 2740.0 + 3390.0] X 10^{6})/ 4 =
2756.0 X 10^{6}

$(2153.8 – 2756.0)^2$ = 362644.8
$(2740.0 – 2756.0)^2$ = 256.0
$(2740.0 – 2756.0)^2$ = 256.0
$(3390.0 – 2756.0)^2$ = 401956.0

$\frac{362644.8 + 256.0 + 256.0 + 401956.0}{4}$ = 191278.2

√191278.2 = 437.4

Percentage error

$\frac{453.6}{3678.4}$ X 100 = 12.3 % $\frac{437.4}{2756.0}$ X 100 = 15.9 %

Combined error

$12.3^2 + 15.9^2$ = combined $error^2$ = 404.1

√404.1 = 20.1 %

$\frac{0.75}{100}$ X 20.1 = 0.15

E.O.P. of CCHalfSλ

2755.95/ 3678.4 = 0.75 ±0.15

MCCHalfS fusion MTase:

Unmodifiedλ Calculations:

1 mL / 100 µL = 10

87 plaques X 10 X dilution = 430 X 10^{-6}

= 430 X 10^{6} undiluted phage

Numbers of phage plaques are adjusted for differences in dilution.

Control (None) average- ([430 + 190 + 510] X 10^{6})/ 3 = 376.6 X 10^{6}

$(430.0 – 376.6)^2 = 2916.0$
$(190.0 – 376.6)^2 = 34819.6$
$(510.0 – 376.6)^2 = 17795.6$

$$\frac{2916.0 + 34819.6 + 17795.6}{3} = 18510.4$$

$\sqrt{18510.4} = 136.1$

MCHSF MTase + R average- ([57 + 8 + 46] X 10^{6})/ 3 = 37 X 10^{6}

$(57 – 37)^2 = 400$
$(8 – 37)^2 = 841$
$(46 – 37)^2 = 81$

$$\frac{400 + 841 + 81}{3} = 440.6$$

$\sqrt{440.6} = 21.0$

Percentage error

$\frac{136.1}{376.6}$ X 100 = 36.1 % $\frac{21.0}{440.6}$ X 100 = 4.8 %

Combined error

$36.1^2 + 4.8^2$ = combined error^2 = 1326.3

$\sqrt{1326.3} = 36.4$ %

$\frac{0.1}{100}$ X 36.4 = 0.04

E.O.P. of Unmodifiedλ

37/ 376.6 = 0.10 ±0.04

MCHSFλ Calculations:

1 mL / 100 µL = 10

87 plaques X 10 X dilution = 870 X 10^{-6}

= 870 X 10^{6} undiluted phage

MCHSF MTase average- ([870 + 1330 + 1450] X 10^6)/ 3 = 1216.6 X 10^6

$(870.0 – 1216.6)^2 = 120131.6$
$(1330.0 – 1216.6)^2 = 12859.6$
$(1450.0 – 1216.6)^2 = 54475.6$

$\frac{120131.6 + 12859.6 + 54475.6}{3} = 62488.9$

$\sqrt{62488.9} = 250.0$

MCHSF MTase + R average- ([980 + 1420 + 1330] X 10^6)/ 3 = 1243.3 X 10^6

$(980.0 – 1243.3)^2 = 69326.9$
$(1420.0 – 1243.3)^2 = 31222.9$
$(1330.0 – 1243.3)^2 = 7516.9$

$\frac{69326.9 + 31222.9 + 7516.9}{3} = 36022.2$

$\sqrt{36022.2} = 189.8$

Percentage error

$\frac{250.0}{1216.6} \times 100 = 20.5\ \%$ $\frac{189.8}{1243.3} \times 100 = 15.3\ \%$

Combined error

$20.5^2 + 15.3^2 = \text{combined error}^2 = 654.3$

$\sqrt{654.3} = 25.6\ \%$

$\frac{1.02}{100} \times 25.6 = 0.26$

E.O.P. of MCHSFλ

$1243.3/ 1216.6 = 1.02 \pm 0.26$

Appendix F

MS fusion Peptide Fragmentation MS Results:

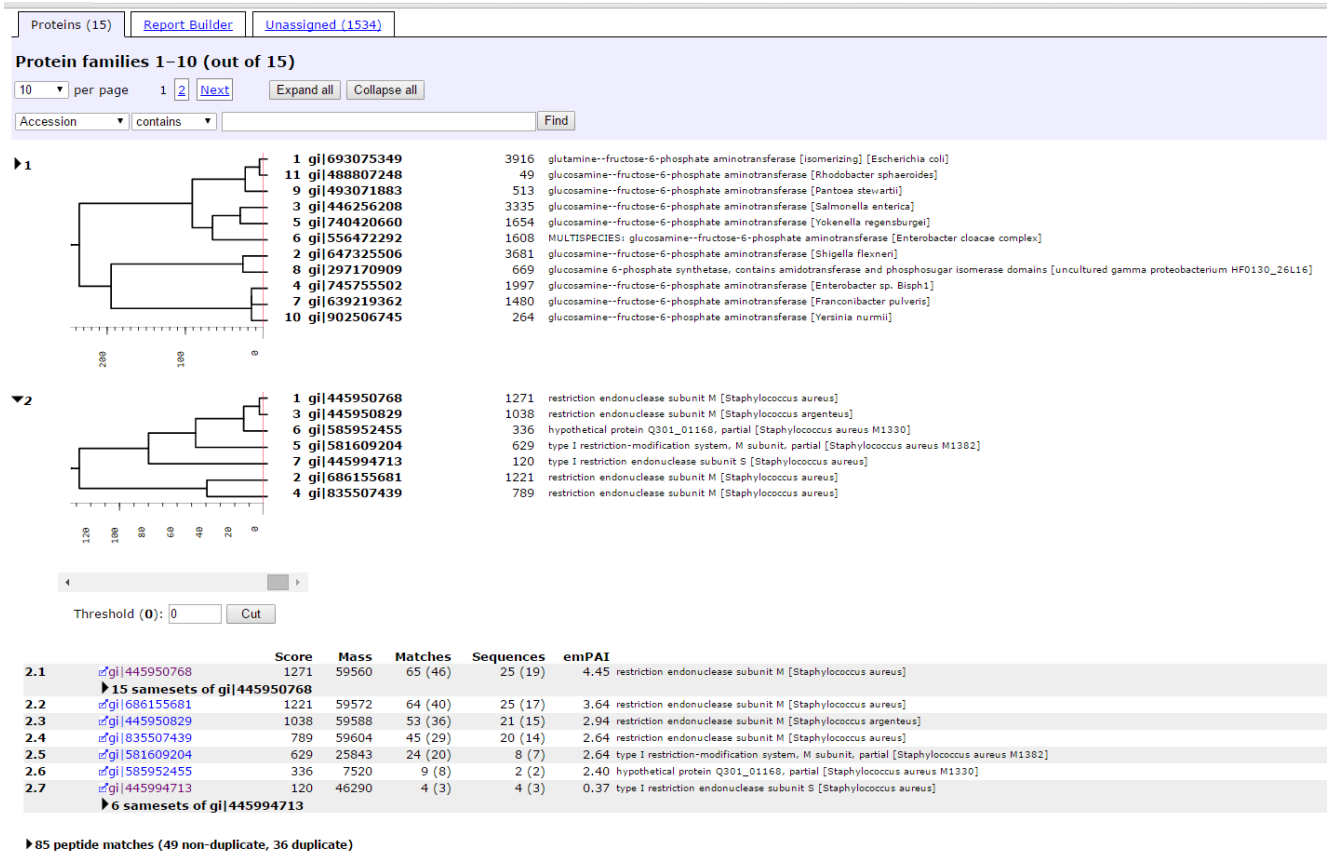


Figure 119: Breakdown of enzyme sequence matches to peptide fragment mass results.

MASCOT Search Results

Protein View: gi|445950768

restriction endonuclease subunit M [Staphylococcus aureus]

Database: NCBInr
Score: 1271
Nominal mass (M_r): 59560
Calculated pI: 4.63
Taxonomy: Staphylococcus aureus

This protein sequence matches the following other entries:

- gi|329313150 from Staphylococcus aureus subsp. aureus T0131
- gi|421957363 from Staphylococcus aureus CN79

Sequence similarity is available as an NCBI BLAST search of gi|445950768 against nr.

Search parameters

MS data file: C:\data from 5600+\151105\Edward 60 kDa 3uL.mgf
Enzyme: Trypsin: cuts C-term side of KR unless next residue is P.
Fixed modifications: Carbamidomethyl (C)
Variable modifications: Oxidation (M)

Protein sequence coverage: 49%

Matched peptides shown in ***bold red***.

```
  1 MSITEKQRQQ QAELHKKLWS IANDLRGNMD ASEFRNYILG LIFYRFLSEK
 51 AEQEYADALS GEDITYQEAW ADEEYREDLK AELIDQVGYF IEPEDLFSAM
101 IREIETQDFD IEHLATAIRK VETSTLGEES ENDFIGLFSD MDLSSTRLGN
151 NVKERTALIS KVMVNLDDLP FVHSDMEIDM LGDAYEFLIG HFAATAGKKA
201 GEFYTPQQVS KILAKIVTDG KDKLRHVYDP TCGSGSLLLR VGKETQVYRY
251 FGQERNNTTY NLARMNMLLH DVRYENFDIR NDDTLENPAF LGNTFDAVIA
301 NPPYSAKWTA DSKFENDERF SGYGKLAPKS KADFAFIQHM VHYLDDEGTM
351 AVVLPHGVLF RGAAEGVIRR YLIEEKNYLE AVIGLPANIF YGTSIPTCIL
401 VFKKCRQQDD NVLFIDASND FEKGKNQNHL SDAQVERIID TYKRKETIDK
451 YSYSATLQEI ADNDYNLNIP RYVDTFEEEA PIDLDQVQQD LKNIDKEIAE
501 IEQEINAYLK ELGVLKDE
```

Figure 120: Peptide mass result matches to the HsdM protein.

MASCOT Search Results

Protein View: gi|445994713

type I restriction endonuclease subunit S [Staphylococcus aureus]

Database: NCBInr
Score: 120
Nominal mass (M_r): 46290
Calculated pI: 9.36
Taxonomy: Staphylococcus aureus

Search parameters

MS data file: C:\data from 5600+\151105\Edward 60 kDa 3uL.mgf
Enzyme: Trypsin: cuts C-term side of KR unless next residue is P.
Fixed modifications: Carbamidomethyl (C)
Variable modifications: Oxidation (M)

Protein sequence coverage: 12%

Matched peptides shown in ***bold red***.

```
  1 MSNTQKKNVP ELRFPGFEGE WEEKKLGEFA GKVTQKNVDK KYIETLTNSA
 51 ELGIISQKDY FDKEISNIDN IKKYYVVEEN DFVINPRMSN YAPFGFVNRN
101 KLGKKGVMSP LYTVFKIQNI DLNFIEFYFK SSKWYRFMAL NGDSGARADR
151 FSIKDRTFME MPLHIPCMDE QIKIGQFFSK LDRQIELEEQ KLELLQQQKK
201 GYMQKIFSQE LRFKDENGKD YPEWEETTIK EIAQINTGKK DTKDAITNGS
251 YDFYVRSPIV YKINTFSYEG EAILTVGDGV GVGKVFHYVN GKFDYHQRVY
301 KISDFKNYYG LLLFYYFSQN FLKETKKYSA KTSVDSVRKD MIANMKVPRP
351 IYIEQKKIGQ FIKRVDNKTK IQKQVIELLK QRKKSLLQKM FI
```

Figure 121: Peptide mass result matches to the HsdS protein.

Appendix G

HalfSHis SMRT Results:

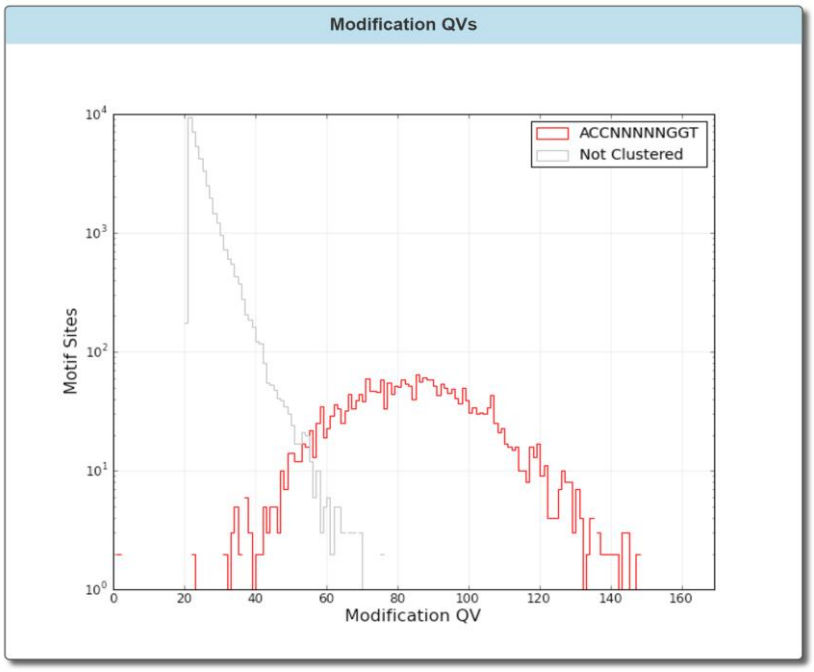


Figure 122: DNA sequence motif detection (B). The graph shows DNA sequence motif detection against detection of methylated motifs. ACCN$_5$GGT (Red) is the only detected motif.

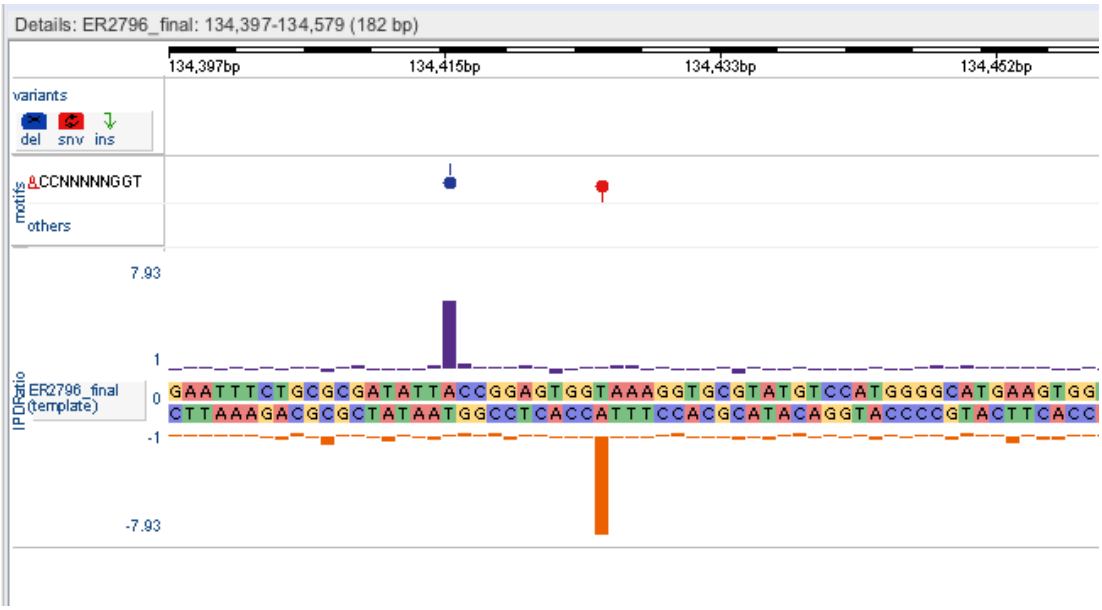


Figure 123: Example of a detected motif. The bars are proportional to the time taken for the nucleotide to be incorporated into the growing chain. The high bars on the top and bottom adenines indicate that it was a relatively longer period to add these bases, and they are therefore recognised as being methylated.

CCHalfSHis SMRT Results:

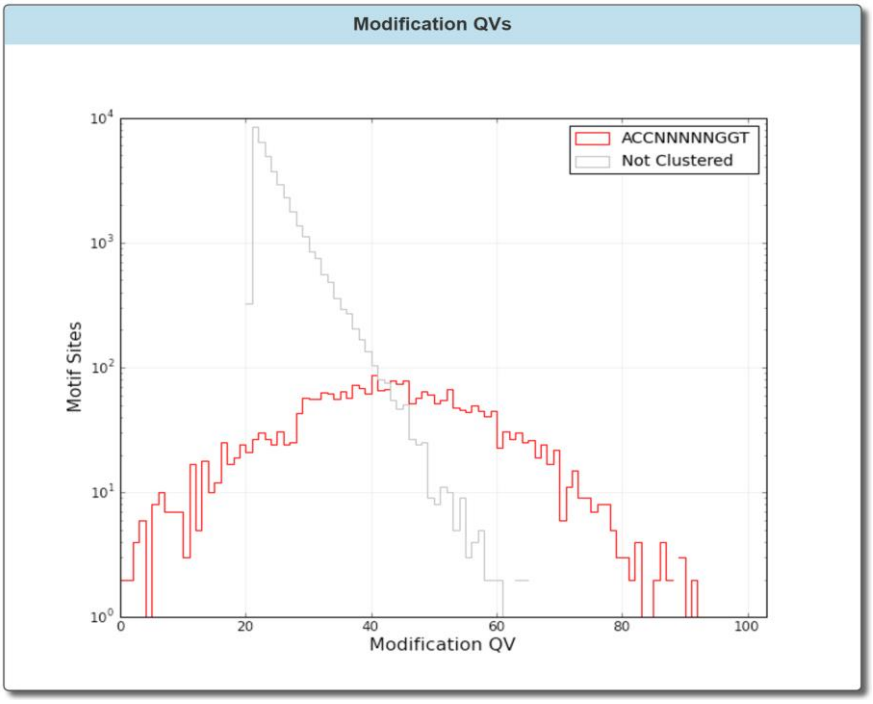


Figure 124: DNA sequence motif detection (B). The graph shows DNA sequence motif detection against detection of methylated motifs. ACCN$_5$GGT (Red) is the only detected motif.

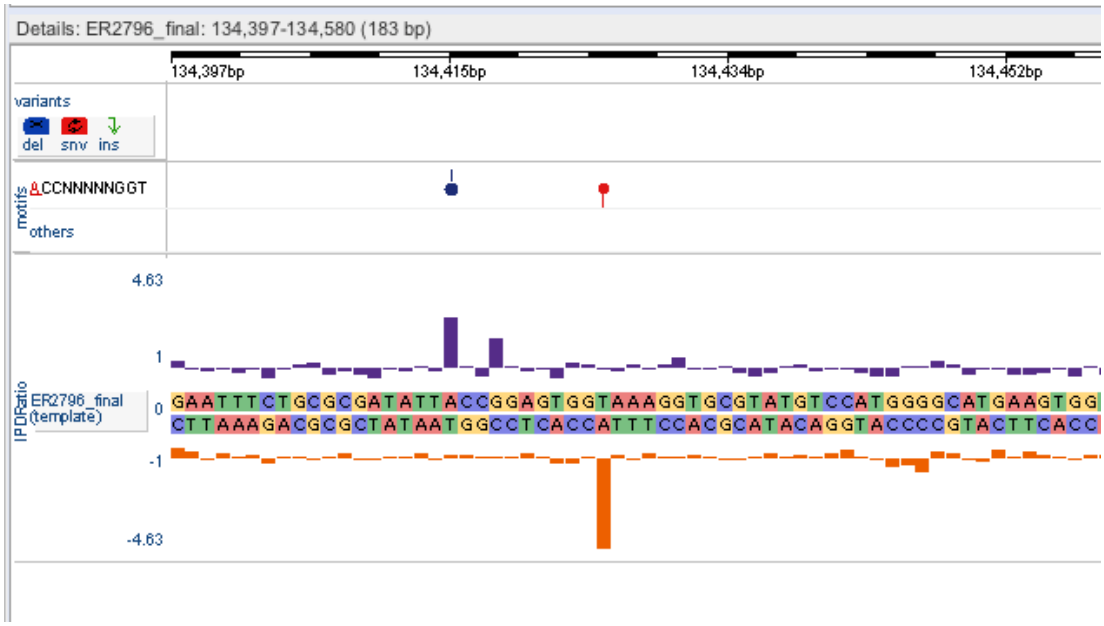


Figure 125: Example of a detected motif. The bars are proportional to the time taken for the nucleotide to be incorporated into the growing chain. The high bars on the top and bottom adenines indicate that it was a relatively longer period to add these bases, and they are therefore recognised as being methylated.

MCCHalfS fusion SMRT Results:

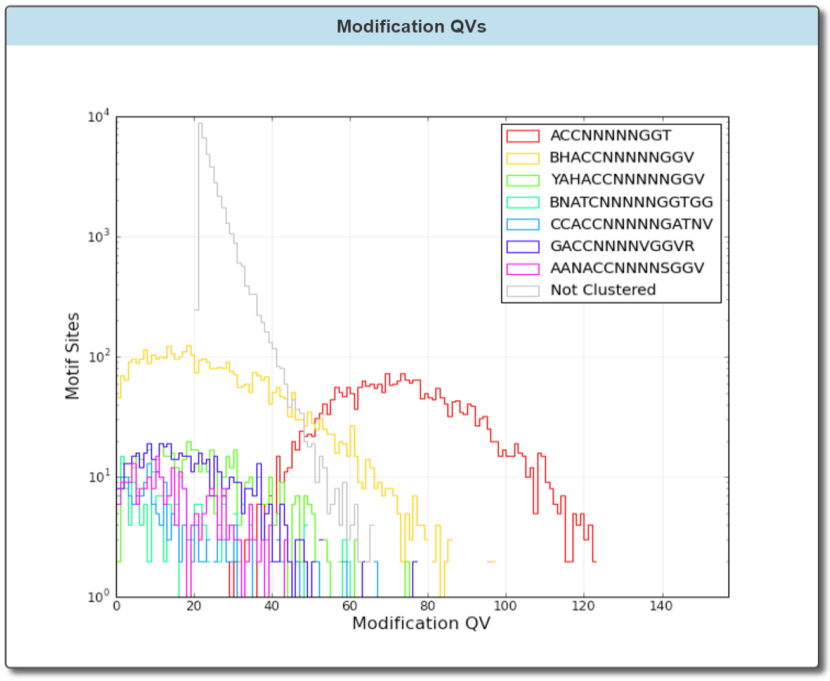


Figure 126: DNA sequence motif detection (B). The graph shows DNA sequence motif detection against detection of methylated motifs. ACCN$_5$GGT (Red) is the most frequently detected motif

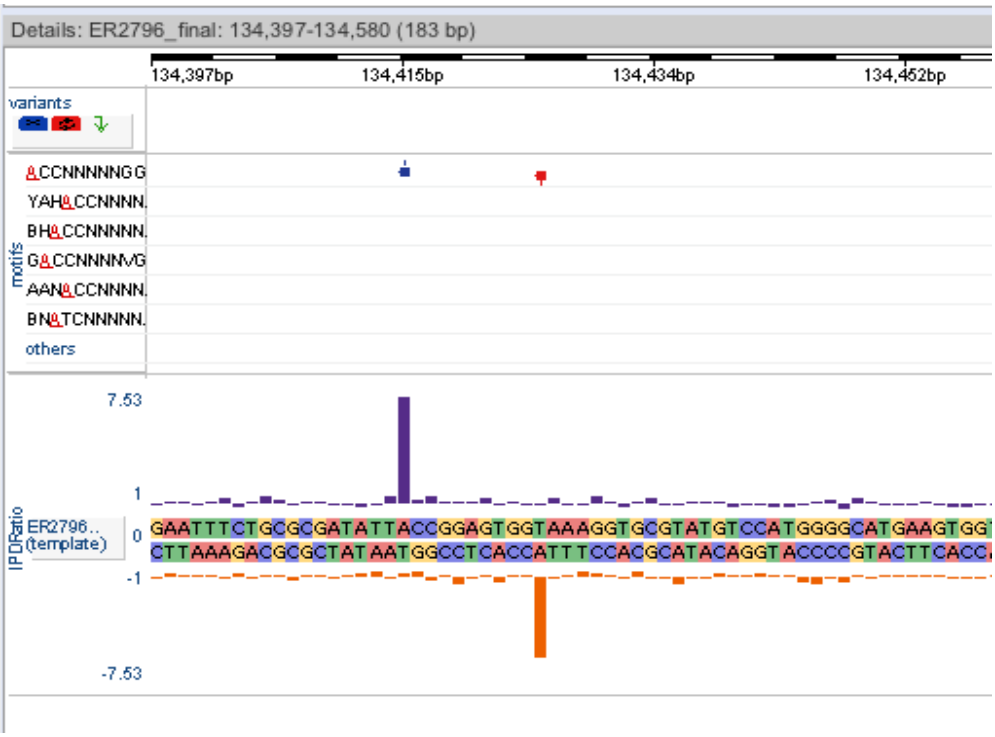


Figure 127: Example of a detected motif. The bars are proportional to the time taken for the nucleotide to be incorporated into the growing chain. The high bars on the top and bottom adenines indicate that it was a relatively longer period to add these bases, and they are therefore recognised as being methylated.